The Impact of Second and Fifth Grade Teacher Experience and Effectiveness on Student Achievement in English Language Arts and Mathematics Measured by the Michigan Education Assessment Program

Tammy DiPonio
B.A., California University of Pennsylvania, 1993
M.S., Northwest Missouri State University, 1998

Submitted to the Graduate Department and Faculty
of the School of Education of Baker University
in partial fulfillment of the requirements for the degree of
Doctor of Education in Educational Leadership

_____
Sharon Zoellner, Ph.D., Major Advisor

_____
Dennis King, Ed.D.

_____
Stephani Reynolds, Ed.D.

Date Defended: April 4, 2016

Abstract

Determining how to accurately measure the impact teachers have on student achievement is an ongoing topic in education. The purpose of this study was to determine if teacher performance, as measured by teachers' final evaluation ratings, had an impact on elementary students' academic achievement. The second purpose of the study was to determine if teachers' years of experience made an impact on students' academic performance. The study also investigated if there was an interaction between teacher experience and teacher effectiveness. A non-experimental, ex-post facto research design was used for this comparative data study. The independent factors were the teacher evaluation rating and years of experience. The dependent variables were the average classroom scores of the Michigan Educational Assessment Program (MEAP) in English language arts and mathematics. The population included a group of second and fifth-grade teachers rated highly effective and effective. Two years of data were analyzed. A 2 x 3 factorial ANOVA was used to compare student scores on the MEAP by teacher effectiveness ratings and years of experience. Results revealed that second-grade and fifth-grade teachers' performance and experience had no statistically significant findings based on third-grade and sixth-grade students' mathematics MEAP performance. The third-grade English language arts scores indicated that second-grade teachers' performance and experience had no statistically significant findings. The English language arts scores for sixth-grade differed for teacher effectiveness ratings. Teachers who were rated effective had higher class mean scores than teachers who were rated as highly effective. No statistically significant differences were found for years of experience or the interaction between teacher effectiveness and years of experience. This

study could be used by states and districts to influence policies for teacher evaluation, including student growth targets, and professional goals. Future replication of this study could include additional grade levels, studying more than two districts, and including districts with varied socioeconomics. By expanding the study, results may reveal if student performance is influenced by teacher effectiveness ratings.

Dedication

This dissertation is dedicated to my family. To my mom, continually encouraging me to keep going, always there to listen as I shared the obstacles I encountered along the way. To my dad, thank you for modeling and instilling in me the values of hard work, effort, and dedication. Without the perseverance that you taught me, I would not have been able to accomplish this achievement. To Dominick, Natalie, and Nikko, you put up with all of my hours in front of the computer, headphones on as I tuned out the world around me. I am looking forward to never again missing out on your events and activities. I might even be ready for another trip to Tawas! Joe, as I spent endless hours behind the computer screen, or tucked away at the library you took on all of the "other" responsibilities by yourself, I appreciate all of your help along the way. Thank you for reading my edits, answering my questions, and reassuring me. To Terry, Rob, and Deb you always asked how it was going and kept encouraging me. I love and appreciate all of you and appreciate your patience, understanding and your unending support. I am thankful to God for helping me get through this. I look forward to having time to serve God by doing more for others.

Acknowledgements

Though too many to mention, I have been fortunate to work with so many outstanding colleagues and mentors throughout my career. Through our relationships you have inspired me, challenged me, and encouraged me. I will be forever grateful for the opportunity to learn and grow through our work together. Though it has been a while since I have been in class, the application of my doctoral coursework was influential in my administrative career. To my advisors and professors at Baker University, especially Dr. Zoellner, Dr. Hole, and Dr. Messner, I appreciate your time, guidance, and most of all your patience throughout my study. Dr. Zoellner, thank you, for your support and unending patience. Whether it was a Zoom meeting, late night e-mails, or phone call, our meetings helped to clarify the next steps. Dr. Messner, your statistical expertise was greatly appreciated. June, I could not have done this without you. You helped me every step of the way and I am so very grateful to you. To all the carpool moms, I owe you! I can finally reciprocate; just send your kids to spend the summer with me. Concerts, swimming, trips to Birmingham, whatever they want, I am thrilled to be dedicating the next chapter of my life to the teenagers! Finally, Dr. Reynolds, and Dr. King thank you for your willingness to participate on my dissertation committee.

Table of Contents

List of Tables

List of Figures

Chapter One

Introduction

In 1983, American schools were failing according to A Nation at Risk (National Commission on Excellence in Education, 1983). The National Commission on Excellence in Education called for improved teacher preparation, stating, "Teachers need to meet high educational standards, demonstrate an aptitude for teaching and demonstrate competence in an academic discipline" (Caliendo, 1986, p. 22). In 2001, President George Bush proposed the No Child Left Behind Act (NCLB) to address the "great concern that too many of our neediest children are being left behind" (U.S. Department of Education [USDE], 2002, p. 1). A goal of NCLB was to increase the accountability of "highly qualified teachers" (p. 3). In 2009, the U.S. Department of Education federal program known as Race to the Top (RTT), authorized under the American Recovery and Reinvestment Act of 2009 (ARRA), offered funding incentives to encourage states to reform teacher evaluation systems. The focus of new evaluation systems shifted to student achievement results rather than to rely exclusively on teacher qualifications (State of the States, 2011). In addition to attempting to reward schools for improved teacher evaluation programs, RTT was a catalyst for innovative educational programs such as 21st-century community learning centers, migrant education programs, assistive technology, and longitudinal data systems (Devine, 2009). With increased expectations and greater accountability by educators, quality school programming and highly qualified teachers became a leading educational priority. Accountability has become widespread continuing to dominate the literature and criticism of the public schools. One recurring claim is that the tenure system is a safety net for ineffective teachers (Weisberg, Sexton,

Mulhern, & Keeling, 2009). President Obama stated, "I reject an educational system that rewards failure and protects a person from its consequences" (Weisberg et al., 2009, p. 2). According to Weisberg et al., (2009), teacher evaluation systems are superficial, capricious, and often do not directly address the quality of instruction school districts require. "A troubled state of teacher evaluation is a glaring and largely neglected problem of public education, one with consequences that extend far beyond the performance-pay debate" (Toch & Rothman, 2008, p. 1). Researchers blamed failing students on the lack of skilled teachers (Bill & Melinda Gates Foundation, 2013; Toch & Rothman, 2008; USDE, 2009; Weisberg et al., 2009).

Evaluation systems across the country were reviewed in response to extensive criticism. Critics were concerned that 99% of teachers were rated in the top two levels, effective, and highly effective. The term effective is one of four evaluation labels established by the Michigan Department of Education. An effective rating is the second highest level a teacher can earn in the teacher evaluation system, while the highly effective label is the highest level that Michigan teachers can receive for their evaluation. In 2011, the Michigan Council for Educator Effectiveness was created (Nowlin, 2011). Their charge was to recommend a process for reviewing local district evaluation tools and to recommend a state evaluation tool for teachers and administrators to the governor, the state legislature, and the state board. Michigan, like many other states across the nation, became involved in studying, revising, and changing educator evaluations. Student growth, professional learning, and enhanced instruction were top priorities (NCTQ, 2014). As the new evaluation systems evolved, many included teacher ratings on various measures, not just on classroom observations. The multiple measures

encompassed student growth, professional contributions, reflective practice, planning, and observation of instruction (USDE, 2013).

Educators across the state made recommendations for a new state teacher evaluation system. One participant on the Michigan State Task Force of K-12, Jennifer Hammond, Grand Blanc High School Principal, believed the state needed to take ownership in the evaluation system, teacher training, and assessment writing, instead of having 500 districts doing different things (French, 2014). Hammond thought a consistent tool for evaluators that allowed for a more accurate assessment of teacher performance could make a greater impact on student growth and achievement. The legislation required all districts to include student growth as a "significant part" of a teacher's final evaluation.

In 2013-2014, the Michigan Compiled Law (MCL) 380.1249 was created to ensure that student growth would account for 25% of the teacher's overall evaluation ranking. By 2014-2015, 40% of the overall evaluation was based on student growth, and was to increase to 50% by 2015-2016. Schools measured student growth in various ways; through district common assessments, norm-referenced data, work samples, end-of-course exams, and standardized test data (Michigan Department of Education [MDE], 2014). Table 1 displays the weighted percentages for student growth over time.

Table 1

Michigan Legislative Timeline for Percentage of Teacher Evaluation Based on Student Achievement Data

School Year	Tool Type	% of evaluation based on student achievement and achievement data	Reporting Requirement
2011-2012	Locally determined	Significant part	Effectiveness labels in June Registry of Educational Personnel (REP) collection
2012-2013	Locally determined & Michigan Council for Educator Effectiveness (MCEE) Pilot	Significant part	
2013-2014	MCEE's evaluation tool	25%	
2014-2015	MCEE's evaluation tool	40%	
2015-2016	MCEE's evaluation tool	50%	

Note. Adapted from *Understanding Educator Evaluations in Michigan* (Rep.). Michigan Department of Education, 2012. Available at https://www.michigan.gov/documents/mde/Educator_Effectiveness_Ratings_Policy_Brief_403184_7.pdf

Historically, many Michigan districts had evaluation systems that included two ratings, *Satisfactory* or *Unsatisfactory*. As of 2013 there were four rating levels statewide: *Highly Effective*, *Effective*, *Minimally Effective*, or *Ineffective*. Evaluation categories included planning and preparation, student growth, instructional delivery, environment, and professional responsibilities. Some categories were weighted more heavily than others. Some categories might also differ from one district to another. The cut scores for each of the categories dictated the teacher's overall effectiveness ranking. In 2012, Michigan's overall teacher evaluation results indicated the majority of teachers

(98%) were in the two top categories with 23% falling into the highly effective category and 75% scoring in the effective category (Keesler & Howe, 2012). Though state laws require student growth to be the evaluation component with the highest weight, exactly how one district determined an effective teacher rating may have been very different from how another district determined an effective teacher.

Beyond rating teachers, data from teacher evaluations were used to determine what professional development or instructional coaching support a teacher might need (The National Education Association, 2011). Some districts used the data to promote a teacher or assign roles beyond the classroom. A district also could use the growth results for compensation or merit pay. Additionally, removal and termination decisions were made based on the evaluation results.

Background of the Study

The focus of the current study was to determine the impact of teacher evaluation and teacher experience on student achievement for third and sixth-grade students as measured by the Michigan Educational Assessment Program (MEAP). NCLB required schools to be held accountable for student academic achievement. At the time of this study the MEAP was the state accountability test used in Michigan to meet the NCLB requirements. Mandated by the state of Michigan, the district and school results were communicated to the public through Accountability Scorecards, posted on district websites, and through the state department of education. The scorecard measure replaced the adequate yearly progress (AYP) report. Schools and districts earned points toward the scorecard for assessment scores, including MEAP results, graduation rates, and compliance with state and federal laws. The points were calculated into a scorecard

rating. To learn more about a school or district's strengths and weaknesses; parents, prospective community members, and all stakeholders were able to access the Accountability Scorecards through the state and local districts (MDE, 2014).

Table 2 depicts Michigan's 900 school districts, including charter schools that educate 1,529,887 students in kindergarten through grade 12 (Michigan Department of Education, 2014).

Table 2

2012-2014 Public School Districts in Michigan

Type of School District	2012-2013	2013-2014
Intermediate School Districts	56	56
Local Educational Authorities	549	545
Charter Schools	277	298
Total	883	900

Note. Adapted from *Center for Educational Performance* (2015). https://www.mischooldata.org/DistrictSchoolProfiles/ReportCard/EducationDashboard.aspx

Public schools in Michigan were required to comply with the NCLB mandates through the MEAP. Starting in third grade, students took the MEAP assessment each year. The MEAP assessment was aligned with the state standards. Schools were expected to teach the state content standards. The MEAP assessment measured how well schools were performing on mastering the standards (MDE, 2014).

As referenced by the Bureau of Assessment and Accountability (BAA), an assessment division of the Michigan Department of Education, the state MEAP assessments should measure what Michigan educators believe all students should know and be able to achieve in the core content areas of mathematics, English language arts,

and science (MDE, 2011). MEAP assessment results reveal how well students and schools perform in relation to the state standards. The MEAP assessment results offer one perspective on school performance.

The most important factor for schools in improving student achievement is teacher effectiveness, which is not measured, recorded, or used to inform decision-making in any meaningful way (Weisberg et al., 2009). The new structure of teacher evaluations seeks to hold teachers to higher and more universal standards. The purpose of the new system was to make the evaluation process more meaningful, eliminate compliance without purpose, establish a system that provided educators with useful feedback to enhance their development, and include student data to measure teacher performance (Weisberg et al., 2009).

Nationwide, educators recognized the imperfections of teacher evaluation systems. Ineffective teachers were not addressed consistently from one district to another, and few evaluation systems recognized quality teachers (New Teacher Project, 2010). A limited number of teachers fell into the minimally effective and ineffective categories. Stronge wrote, "Research has found, in the typical district: 20 percent of teachers are 'highly effective,' 60 to 65 percent are 'effective,' 10 percent are 'partially effective' and 3 to 5 percent are 'ineffective" (as cited in Mooney, 2012, p. 2). Even though most evaluation systems are marketed as growth models, minimal support is given through the evaluation process to enhance teacher growth and development (Weisberg et al., 2009). These concerns initiated the revision of teacher evaluation systems. Many districts developed evaluation systems that included student growth as a major component of teacher evaluations.

Teacher evaluation systems are being adjusted to address ineffective teachers. At the time of this study, inconsistencies existed among districts when rating teachers based on student achievement. Using student performance as a significant component of a teacher's overall evaluation rating is problematic unless adjustments are made that could account for factors such as students with special needs, demographic factors, and differences in class sizes (Shavelson et al., 2010).

Changes on teacher performance evaluations were being made in education. In Michigan, evidence of student growth and achievement were weighted higher than any other component of the teacher evaluation rubric (Michigan Council for Educational Effectiveness, 2013). This evidence is significant as it affects teachers' overall evaluation ratings. Teachers who receive the highest rating also receive an additional stipend through performance-based compensation. State policymakers needed to determine consistent multiple measure assessments to be used by all schools to measure student growth. The lack of consistency that existed meant teachers in one district would be able to achieve a higher evaluation rating than teachers in another district who might have more rigorous goals, making it more challenging to meet. For teachers to improve instruction and positively influence student achievement, administrators must provide quality feedback and support teacher growth through differentiated professional development focused on quality instruction (Michigan Council for Educational Effectiveness, 2013).

This study was focused on two school districts outside of Detroit, Michigan. District A, a suburban district, serves 3,400 students in one high school, one alternative high school, one middle school, and three elementary schools. The two high schools, one

alternative high school, four middle schools, and 12 elementary schools in District B serve approximately 12,000 students. District B is the second most diverse district in Michigan where over 80 languages are spoken (MDE, 2014).

Statement of Problem

For the state of Michigan, no published literature has been found that supports a link between student achievement and teacher effectiveness ratings. However, in Michigan schools, teacher effectiveness ratings are influenced by student academic growth. Regardless of the student test scores, the majority of teachers are rated either highly effective or effective. Few teachers are rated minimally effective or ineffective, yet students repeatedly fail to reach proficiency levels on standardized tests. Adding additional pressure, the state of Michigan has established performance-based pay incentives for teachers who earn a highly effective rating. However, concerns arise when different schools and different teachers within a school determine student growth in various ways. Since student growth is weighted heavily in the overall evaluation, the districts in the study are investigating consistent ways to measure student growth. Using standardized testing data is a consideration. If valid, using the same test data from all schools within a district and within the state could be one step toward minimizing subjectivity. The intent of this study is to fill a gap in the research regarding teacher effectiveness and academic growth.

Purpose of Study

The purpose of this study was to determine if teacher performance, as measured by teachers' final evaluation ratings, had an impact on elementary students' academic achievement. The second purpose of the study was to determine if teachers' years of

experience had an impact on students' academic achievement. Additionally, the study will determine if there is an interaction between teacher experience and teacher effectiveness, and its impact on student achievement. Specifically, the intent of the current study was to identify if second and fifth-grade teachers' evaluation ratings and years of experience made an impact on students' achievement as measured by the third and sixth-grade fall MEAP assessment for mathematics and English language arts.

Significance of the Study

Teacher evaluation reforms need to assess the connection between evaluation, student achievement, and years of experience (Garnett, 2013). "Connecting effective teacher practice and increased levels of student achievement can further justify the cause of public education" (Alleman, 2006, p. 9). The results of this study could influence those at the state level to further investigate evaluation tools. It is possible that a pattern of inconsistencies will be found. Should the study reveal these inconsistencies, the study would support the need for further analyses of the various evaluation tools. Ultimately, the implementation of a statewide evaluation tool with consistent student growth measures could be recommended to the State Board of Education.

Delimitations

"Delimitations are self-imposed boundaries set by the researcher on the purpose and scope of the study" (Lunenburg & Irby, 2008, p. 134). Delimitations may impact the generalizability of the findings of the study. The following were delimitations of this study:

1. The study involved two school districts.
2. The study used two years of data.

3. The study used quantitative data.
4. The study included students who took the MEAP and were in third-grade and sixth-grade in two Michigan school districts for 2012-2013 and 2013-2014.
5. The study was limited to second and fifth-grade teachers whose annual evaluations were either highly effective or effective.
6. Teacher effectiveness ratings as they pertain to student academic achievement were used. Other portions of the effectiveness ratings were not included in this study.
7. The population of this study did not include teachers rated minimally effective and ineffective due to the limited number of teachers scoring in these categories.

Assumptions

"The assumptions are items taken for granted relative to the study" (Roberts, 2004, p. 129). The following assumptions were made in this study:

1. All students gave their best effort on the MEAP assessments.
2. Data collected from districts was accurately recorded.
3. Administrators were proficient and consistent in the use of the evaluation tools.
4. The final teacher evaluation rating was an accurate reflection of the teacher's effectiveness.
5. The data meet the assumptions of parametric testing; the dependent variable (MEAP scores) is normally distributed and interval scaled and the independent variables are categorical (teacher evaluations – highly effective

and effective and years of experience – new, level I and level II).

Research Questions

The study was conducted using two years of students' fall MEAP data for the 2012-2013 and 2013-2014 school years. The teacher evaluation scores were from the spring teacher evaluation for 2013 and 2014. The students' fall test scores were linked to the teacher who taught the content the previous school year. The independent variables were teacher performance and teacher experience. The dependent variables were the student achievement scores as measured with the fall MEAP mathematics and English language arts assessments. Student achievement data for third grade and sixth-grade was used in the quantitative analysis of this study. The following research questions guided this study:

RQ1. Does second-grade teacher experience, teacher effectiveness, and the interaction of experience and effectiveness influence student mathematics achievement for third-grade students when measured with the fall MEAP?

RQ2. Does second-grade teacher experience, teacher effectiveness, and the interaction of experience and effectiveness influence student English language arts achievement for third-grade students when measured with the fall MEAP?

RQ3. Does fifth-grade teacher experience, teacher effectiveness, and the interaction of experience and effectiveness influence student mathematics achievement for sixth-grade students when measured with the fall MEAP?

RQ4. Does fifth-grade teacher experience, teacher effectiveness, and the interaction of experience and effectiveness influence student English language arts achievement for sixth-grade students when measured with the MEAP?

Definition of Terms

The following section contains terms that were used throughout the study. The terms were explicitly defined to ensure the reader had a clear understanding.

Criterion-referenced. Criterion-referenced tests measure a student's achievement against curriculum content or established standards. MEAP tests are criterion-referenced tests. Cut scores are determined by the test developers. Test results show whether a student scored above or below the established cut scores (MDE, 2011).

Effective. The effective label is one of four evaluation labels established by the Michigan Department of Education. An effective rating is the second highest evaluation level a teacher can earn (Keesler & Howe, 2012).

Highly qualified teacher. The highly qualified teacher status is a minimum requirement that must be obtained to become a teacher. Requirements include a bachelor's degree and valid state certification with no requirements waived; the teacher cannot have an emergency or conditional certificate. The teacher must demonstrate expertise in the core academic subject(s) they teach (MDE, 2007). Though similar terminology, the highly qualified teacher term is used for state certification and is different from the term *Highly Effective*, as used in some annual teacher evaluation rating models. A highly qualified teacher, through the evaluation model, could be highly effective, effective, minimally effective, or ineffective.

Highly Effective. One of four evaluation ratings established by the Michigan Department of Education. The highly effective label is the highest level that Michigan teachers can earn for their evaluation rating (Keesler & Howe, 2012).

Level I Experienced Teacher. The label was created by the researcher to

categorize teachers in their fourth through twelfth year of teaching.

Level II Experienced Teacher. The label was created by the researcher to categorize teachers who have been teaching more than twelve years.

Mean Classroom Score. The mean is the arithmetic average (Johnson & Christensen, 2008). In this study, the mean classroom score will be determined by averaging the academic achievement scaled scores for all of the students whose teachers were involved the study. Both the MEAP English language arts and the MEAP mathematics assessment data will be used.

New Teacher. The label used when a teacher is within his or her first three years of teaching and is receiving support from an experienced mentor teacher (Keesler & Howe, 2012).

No Child Left Behind. No Child Left Behind is an Act of Congress passed in 2001 to reform education and improve student achievement in American Schools. Assessments are used to determine if students are mastering standards (USDE, 2002).

Scaled Scores. According to the MEAP Technical Report (2011-2012), "Scaled scores are statistical conversions of raw scores that adjust for slight differences in underlying ability levels at each score point and permit comparison of assessment results across different test administrations within a particular grade and subject" (p. 62).

Overview of the Methodology

A non-experimental research design was used in the study. A 2 x 3 factorial analysis of variance (ANOVA) was conducted to determine if an interaction existed in MEAP mathematics and English language arts outcomes as measured by the average classroom scores for mathematics and English language arts between teachers who were

rated as highly effective or effective and by years of experience. The independent factors were the teacher evaluation rating and years of experience. The dependent variables were the average classroom scores of the MEAP mathematics and English language arts tests.

MEAP tests were given in the fall to students in grades three through nine to measure mastery of state standards in a curriculum area at specific grade levels. The state compiled testing outcomes from districts across Michigan. Individual student, teacher (classroom), school and district scores were available, with school, and district scores communicated to the public through the Michigan Department of Education website (MDE, BAA & Measurement Incorporated, 2011-2012). For the purpose of this study, student achievement data for grades three and six were examined for the mathematics and English language arts MEAP assessments for the 2012-13 and 2013-14 academic years. For example, the MEAP assessment was administered in the fall of 2013 the test was measuring what the student learned in fifth-grade in 2012 therefore, the scores of the sixth-grade tests were linked to the fifth-grade teacher.

Teachers' years of experience were divided into three groups for this study. Table 3 shows the experience labels used for teacher experience.

Table 3

Years of Experience for Study

Experience Label	Number of Years
New Teacher	1-3 years
*Level I Teacher	4-12 years
*Level II Teacher	Exceeds 12 years

Note. *Labels were determined by the researcher

The experience levels were new teacher, level I, or level II. A teacher in any of the three experience levels could be rated using one of the four rating levels. Due to the small number of teachers included as either minimally effective or ineffective, the researcher used only teachers who were rated highly effective and effective. The analysis examines the differences in student achievement scores in mathematics and English language arts by teachers' years of experience, teachers' ratings, and the interaction between teacher experience and teacher ratings (MDE, 2014).

The archival data used in this study were retrieved from two suburban school districts in Michigan. Test data were used for all students, in grades three and six, who took the MEAP assessment in 2012-2013 and 2013-2014. The student scaled scores for each teacher were averaged to obtain a mean classroom score for mathematics and English language arts. Teacher evaluation ratings and years of experience for second and fifth-grade certified classroom teachers were retrieved from each district's data warehouse. The teacher years of experience will be categorized into three groups based on experience.

Organization of Study

The introduction of the study, background and conceptual framework, problem statement, purpose, significance of the study, delimitations, assumptions, research questions, the definition of terms, and the overview of the methodology were presented in chapter one. Chapter two contains the review of literature, which includes relevant research of teacher performance ratings and their connection to student achievement. The research design and methodology are described in chapter three. The results of the data analysis and hypothesis testing are presented in chapter four. The summary of the

findings related to the literature and the implications for action and recommendations for future research are included in chapter five.

Chapter Two

Review of Literature

Across the nation public schools are held accountable for student performance on standardized tests, the Race to the Top education initiative, through President Obama's administration, motivated districts with federal dollars to revamp teacher evaluation systems and to include and align student performance in annual evaluations (Toppo, 2013). This literature review contains research related to teacher effectiveness and its impact on student achievement. Throughout the country as accountability increases, teacher evaluation systems are being revised to ensure that student growth measures are a significant portion of annual teacher evaluations. Topics addressed include the history of standardized testing, history of evaluation systems, increased accountability, Michigan Educational Assessment Program (MEAP) state assessment instrumentation, the Danielson Framework for Teaching evaluation tool, accountability through supervision, practices in teacher evaluation, specific models of teacher evaluation systems, districts' evaluation processes, the teacher evaluation process, studies of teacher performance and years of experience and student achievement, and improving teacher evaluation.

History of Standardized Testing

The United States military began using standardized tests in 1914 for placement purposes. Robert Yerkes, a psychologist, worked for the army and navy from 1924-1944 (Murchison, 1930). Yerkes developed intelligence tests for recruits (Sokal, 1987). The Armed Forces of the United States assessed all potential recruits with the Armed Services Vocational Aptitude Battery (ASVAB) for purposes of determining qualifications for enlistment into the military (Wigdor & Green, 1991).

Since the early 1900s, The American College Testing (ACT) and Scholastic Aptitude Test (SAT) have been used for college and university admission (ACT, 2014; Stickler, 2007). In the 1930s public schools used test scores to select the candidates who would receive scholarships. Colleges continue to use the ACT and SAT to compare student academic performance among students from different high schools. The test scores help universities determine the appropriate course placement options. Students who achieved the individual ACT Benchmarks were more likely than those who did not meet the benchmarks to succeed in college and to earn a degree in a timely manner (Radunzel & Noble, 2012).

Mandatory testing was in place long before NCLB, for students in elementary, middle, and high school (Barnett, Justice, & Sheridan et al., 2012). As a result of NCLB, all students attending public schools in grades three through eight completed required standardized testing one time per year. As accountability through standardized testing continued test scores tended to take on more significance than thorough understanding and learning through critical thinking (Eisner, 2004). Using elementary student standardized test results to determine teacher effectiveness, school status, and overall student success placed increased emphasis on scores (Popham, 1999).

Over the years, increased focus on standardized testing has evolved in public education. For a variety of reasons, public school educators struggled with the focus on scores from standardized testing (Kohn, 2000). Schools were ranked within the state based on results of state assessments. Instead of focusing on higher-order thinking skills, the state assessments tended to have more multiple-choice, low-level questions. It was easier and less costly to score the multiple-choice tests than open-ended items.

Michigan Education Assessment Program (MEAP) Instrumentation.

The two types of standardized tests are norm-referenced or criterion-referenced. A norm-referenced test classifies students by their achievement level on the test. Based on these scores students can be instructed in their areas of strength and weakness. These tests also can be used for class placement in advanced or remedial courses. A criterion-referenced test is a test that measures a student's performance against curriculum content or established standards. On criterion-referenced tests, students either score above or below a cut score, which is established by an individual school or organization. Initially, the MEAP was a norm-referenced test, but in 1973, it was changed to a criterion-referenced test to measure student achievement against state curriculum standards. This change allowed curricula and teaching methods to be monitored and adjusted (MDE, 2011).

Accountability for student learning must move beyond establishing an environment that is conducive to learning, to showing evidence of student understanding through data (Earl, 2013). Accountability for state standards in Michigan was measured through the Michigan Education Assessment Program (MEAP) assessment. Items for the MEAP test were written specifically to match the Michigan content standards for each grade level. Test items for the MEAP were written by Michigan educators and other educational consultants, in addition to the Bureau of Assessment and Accountability (BAA) test development experts. The committee worked to ensure there was alignment between test items and state standards. The Office of Educational Assessment and Accountability (OEAA) embedded field tested items within the state reported test questions. These issues and answers were reviewed by the Bias/Sensitivity Committee to

determine if they would be used for operational assessments, or if they needed to be revised or rejected (MDE, Office of Assessment & Accountability, 2004-2005). MEAP assessment items determined to be unfair, inappropriate, or too difficult could be eliminated (MDE, Office of Assessment & Accountability, 2004-2005).

With input from Michigan educators, the MEAP assessments were produced by the Office of Standards and Assessment (OSA). The assessments were developed in association with guidelines from the federal legislation of NCLB, the USDE, and the MDE guidelines. Tests were distributed through the BAA. The district test coordinators trained the teacher leaders throughout the districts for the administration of the test. These trainers then trained the teachers who administered the test. Yearly, the district followed the MDE guidelines for conducting the test during the October testing window. Upon completion of the testing, all materials were returned to the district MEAP coordinator who then sent the tests to be scored by Pearson Educational Measurement scoring services. The MEAP tests were scored through the Measurement Incorporated scoring center and by trained scorers in Michigan. The constructed response items were sent to the Measurement Incorporated scoring center in Durham, North Carolina (MDE, 2012). Once scoring was completed, all scores were sent to the MDE where scores were compiled for each student.

The MEAP test has been revised continuously to reflect current Grade Level Content Expectations (GLCE) and to eliminate invalid test items. The percentage of students who accurately answered an item, as well as the percent of students who chose the "distracters" on multiple-choice items was calculated. If less than 30% of the students selected the correct answer on a multiple-choice item, the committee reviewed

the question, answer choices, and graphics related to the question to determine what revisions needed to be made for the question to be considered valid (Office of Assessment and Accountability, 2005). The constructed response items were reviewed when no one received the top score. When discrepancies occurred, the test writing staff would analyze the question for flaws, or they would seek additional training on scoring those particular items (Office of Assessment and Accountability, 2005). MDE had a limited number of released items that could be accessed by teachers and parents. The released items were eliminated from any current MEAP tests and were not used in future MEAP assessments. Similarly, the process of evaluating teachers has changed over time. The next section will detail the history of teacher evaluation systems.

History of Evaluation Systems

Teacher evaluation is not a new trend (Matzat, n.d.). In the 20th century, an emphasis was placed on teacher accountability for student academic progress. Teacher evaluation became a priority for schools. However, determining a meaningful approach to support teacher growth and development continues to be a challenge facing schools throughout the country (Matzat, n.d.).

The use of teacher evaluation scales dated back as early as the 1930s (Medley & Coker, 1987). In the 1940s and 1950s appearance, tone of voice, emotional stability, trustworthiness, enthusiasm, and warmth were highly valued teacher traits (Danielson & McGreal, 2000). These characteristics were used as criteria for teacher evaluations. At the time, more studies were being conducted to compare teacher actions and student performance. Researchers did not have studies to validate whether student learning could be linked to the highly valued teacher traits. This lack of research led to further studies

and the development of the clinical supervision model of teacher evaluation (Brophy, 1986).

Research on the history of evaluation of teachers dating back to Post-World War II revealed that administrators needed a more objective tool for teacher appraisals. A clinical supervision model of teacher evaluation became popular in the mid-1950s. In 1966, Cogan modeled the approach with student teachers at Harvard University. The process involved observing with a purpose and following-up with a discussion of growth and reflection on effectiveness. This model provided teachers with a more active role in the supervisory process. Teacher improvement plans were developed through reflection. Plans for improvement of instructional competencies were outcomes of the post-observation discussions (Cogan, 1973).

The formative model emphasized teachers and supervisors working together to identify quality instruction to improve teaching and learning. The state of California was one of the first states to enact formal legislation requiring schools to evaluate students. The Stull Act Assembly Bill 273 (1971), written by Republican John Stull, required certified educators to be formally evaluated every two years using a formalized system established by the local school district. Additionally, the bill required that the teacher evaluation would include assessment of student performance as part of the evaluation process. Tying the analysis of students' academic performance to the evaluation process was not enforced by many schools throughout California. Districts claimed that the lack of clarity prevented them from carrying out this step of the process (Fiorina, 1989). Other criteria were used to determine if a teacher was effective. Some examples included being able to control the class, create a quality learning environment, show competence

in teacher standards, create expectations for student progress, and create ways to check for student understanding. Legislators long before the No Child Left Behind Act of 2001, and the Race to the Top in 2009 had valued teacher accountability. However, the federal accountability was increasingly difficult to impose. In 2015, teacher accountability through standardized test scores resurfaced in evaluations and teachers earning high evaluation ratings were rewarded through performance-based compensation measures (Porter, 2015).

In 1971, Stufflebeam founded the National Joint Committee on Standards for Educational Evaluation (Stufflebeam, 1998). Through his leadership, he brought teachers, administrators, school board members, researchers, and assessment specialists together to study and improve the quality of teaching through the development of standards. Ten years after the committee started, the Standards for Evaluation of Educational Programs were published. Stufflebeam believed teacher evaluation should be directly influenced by the standards. The standards were used for multiple purposes including hiring, retaining, dismissing, planning for professional development, tenure, promotion, merit pay, and remediation. Policymakers and practitioners were encouraged to use the standards to provide direction as they developed teacher evaluation systems. Reference to the standards could dismiss public scrutiny that schools had weak standards and lacked rigor while raising the level of professionalism (Stufflebeam, 1998).

In response to frustrated teachers and principals regarding the evaluation process, new assessments were created for national certification in 1987. The National Board for Professional Teaching Standards created core standards that helped identify quality teachers. The standards focused on five core components: teacher commitment to

student learning, teachers' knowledge of subjects and content, monitoring of student learning, teacher reflection, and teachers' participating as members of a learning community (National Board for Professional Teaching Standards, 2014). Teachers interested in Board Certification participated in rigorous performance-based assessments, including videotaping lessons, analysis of teaching, and artifacts that demonstrated student growth impacted by instruction (Weiss & Weiss, 1998). At the time of the Weiss and Weiss study, evaluation standards recommended by the National Board of Professional Teaching Standards in 1987 were still being used.

Leading up to the 1990s, political and legal interest in the evaluation process continued to press school districts to establish or revise policies on the regulation of evaluations. A greater emphasis was placed on instructional improvement through teacher self-reflection (Robinson, Lloyd, & Rowe, 2008). Though variances existed between school districts on evaluation policies, most addressed the purpose of evaluation, the expected frequency of evaluations, standards for educator performance, and process for determining teacher incompetency. The intent of the evaluation model was to define teacher responsibilities by setting clear outcomes. Recognizing that teachers have different ways of delivering their content, evaluators needed to understand that all teachers would have autonomy in presenting their content. The standards were left to the individual teacher to interpret and guide their instruction.

With the continuation of educational reform in the 1990s, academic standards in public schools were commonly measured through high-stakes tests (Supovitz, 2015). Additional research regarding teacher effects on student achievement emerged. Educational researchers, Tucker and Stronge (2005), explored the connection between

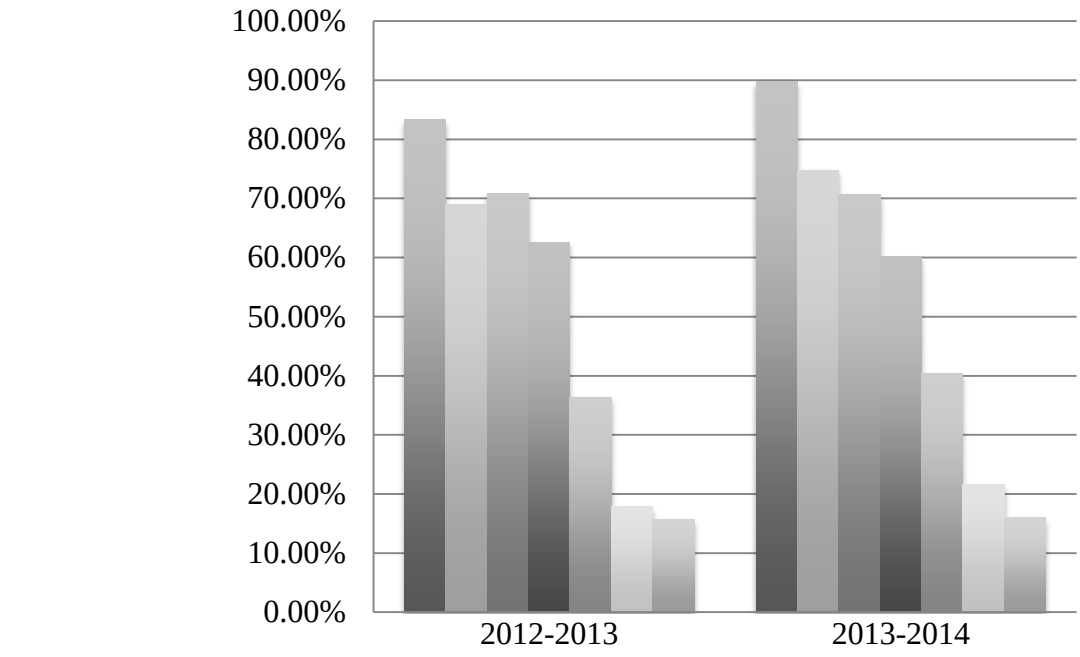
teacher performance and student achievement. A relationship was found between effective teachers and student achievement. "[A] string of highly effective or ineffective teachers will have an enormous impact on a child's learning trajectory" (Palardy & Rumberger, 2008, p. 127). Beyond delivery of instruction, teachers are responsible for learning outcomes. Tucker and Stronge (2005) claimed that effective teachers are the most important factor that influence student learning.

Multiple reports indicated that teacher evaluation systems did not accurately report teacher effectiveness or ineffectiveness. Teacher skill development was lacking because evaluation systems were not addressing these concerns (Bill and Melinda Gates Foundation, 2011; Toch & Rothman, 2008; Weisberg et al., 2009). In response to legislative demands, districts across the United States began reviewing their evaluation procedures. Many states, including Michigan, passed legislation requiring student growth and achievement data to be used as a significant component of the teacher evaluation score.

Teacher Evaluation Systems

"School districts must decide the main purpose of its teacher evaluation system and then match the process to the purpose" (Wise et al., 1984, p. 70). Teacher evaluation systems looked different from district to district, yet regardless of the tool used, the purpose for evaluation system was to improve teacher practices and document a teacher's professional responsibilities and achievements (Marzano, 2012). At the time of the current study, the most significant purpose of evaluations, according to legislative guidelines, was the focus on student growth and achievement (Keesler & Howe, 2012).

Figure 1 compares how Michigan teacher evaluations were used in 2012-2013 and 2013-2014.



	2012-2013	2013-2014
Targeted Professonal Development	83.40%	89.70%
Coaching	69.00%	74.80%
Individualized Development Plan	70.80%	70.70%
Removal or Termination	62.60%	60.10%
School Improvement Plan	36.40%	40.30%
Additional Compensation	17.80%	21.60%
Induction Support	15.70%	16.00%

Figure 1: Decisions Informed by Evaluation Results 2012-2013 and 2013-2014

Note. Adapted from *Michigan Educator Evaluations & Effectiveness Report* by *(MDE),* 2013-2014, Retrieved from https://www.michigan.gov/documents/ mde/2013-14_Educator_Evaluations_and_Effectiveness_ 485909_7.pdf

The Widget Effect was an extensive study of the teacher evaluation process. Approximately 15,000 teachers and 1,300 administrators were part of the study. In the study teacher evaluation processes were analyzed in 12 districts. States involved included Arkansas, Colorado, Illinois, and Ohio. In the Widget Effect, Weisberg, Sexton, Mulhern, and Keeling (2009) reported that quality teachers had a positive impact on student achievement and success, but there are too many teachers rated at the very top level. "Evaluation systems fail to differentiate performance among teachers." (Weisberg et al., 2009, p. 6). The lack of differentiated professional development for teachers was a focus in the Widget Effect. Whether highly effective or effective, minimally effective or ineffective, teachers rarely were given specialized professional development to address strengths or weaknesses (Weisberg et al., 2009). The study outcomes across these districts and states were similar:

- Over 99% of teachers received a satisfactory rating on their teacher evaluations.
- Fifty-nine percent of teachers and 63% of administrators said exceptional teachers were not recognized.
- Professional development was not effectively identified through the evaluation process.
- New teachers were not provided with sufficient support.
- Weak teacher performance was not addressed (p. 6).

Based on the results, observations were infrequent, with some teachers having fewer than two observations in one year. While the length of time for the evaluations varied, all were less than 60 minutes. To increase teacher effectiveness and maximize

student learning, Weisberg et al., (2009) recommended that teacher evaluation systems should involve multiple measures to capture teachers' performance, including student work, evidence of instructional strategies, and evaluation rubrics on behaviors and practice. Multiple observers also should be considered in the teacher evaluation process. When student learning was included as part of the evaluation process and evaluations were not based solely on observation of teaching, teachers earned higher evaluation scores (Toch & Rothman, 2008).

Administrators used the evaluation process to confer with teachers. Administrators offered specific feedback to develop further instructional practices. Using this feedback, teachers could improve and refine their instructional practices. Teachers were given ample opportunities for improvement based on formative evaluation feedback. Principals also used evaluation results to guide teachers toward their professional development needs. Teachers were empowered to self-direct their growth based on feedback from their evaluation (Nolan & Hoover, 2004). Principals offered support and coaching derived from areas in need of improvement. Tenure status and decisions were made based on the evaluation process. According to Nolan and Hoover, teacher effectiveness, depending on the evaluation rubric, was another outcome of the evaluation. Some districts measured and ranked teachers according to the levels on the rubric. In some districts, these rankings were used to reward teachers with performance-based compensation. Districts also used this information if they needed to reduce staff in forced layoffs. While effective evaluation tools are important, the lack of consistency in these tools led to the development of the Danielson's (2007) Framework for Teaching Evaluation Tool.

Danielson Framework for Teaching Evaluation Tool

First publicized in by Danielson, an internationally known expert on teacher effectiveness, the Framework for Teaching is a tool used to measure teacher effectiveness (Danielson, 2007). The Framework for Teaching can be used for all teachers, from elementary classroom teachers to instrumental music instructors to high school biology teachers. The model is designed for use in kindergarten through twelfth grade. The framework includes four domains: (a) planning and preparation; (b) classroom environment; (c) instruction; and (d) professional responsibilities. Each domain is separated into specific components that are narrowed into detailed elements. These components and elements identify clear standards of practice for quality instruction. A rating rubric is used to score the components. During an observation, evaluators collect evidence from each domain based on observed student or teacher behaviors. The elements are scored by the evaluator into one of four performance levels: Highly Effective, Effective, Minimally Effective, and Ineffective. The teacher's overall summative evaluation score is collected through multiple observations. It is recommended that evaluators become certified in the evaluation process through training on observation and scoring (Danielson, 2007).

Research supports the reliability of the Danielson Framework for Teaching (FFT) model (Aramath, 2014; Goe et al., 2008; Kane & Staiger, 2010; Milanowski, 2011). In 2013, the Measures of Effective Teaching (MET) project, funded by the Bill and Melinda Gates Foundation, released its third report on the findings of effective teaching. Multiple research teams evaluated 3,000 teachers who volunteered to be evaluated using the Danielson FFT. Findings indicated that teachers with high observation ratings on the

four domains within the Danielson FFT had high student scores on the standardized assessments. According to the MET study, students scored higher on standard or alternative assessments than on state assessments. Multiple observations resulted in higher reliability. When multiple evaluators observed a different lesson by the same teacher, there was less variation in the overall evaluation score. Through the MET study, researchers found that the teacher is the single most important factor in contributing to student achievement. Veteran teachers became more effective in closing the achievement gap during the year they participated in evaluations using the Danielson FFT (Danielson, 1996).

Sartain, Stoelinga, and Brown (2011) conducted a study of the reliability and validity of the Danielson FFT as they conducted a two-year study on teacher evaluation in the Chicago Public Schools. For the Excellence in Teaching Study (Sartain et al., 2011) and to ensure inter-rater reliability, two evaluators rated the same lesson. Evaluators included a building administrator and an external evaluator. The external evaluator tended to score teachers more critically than the building administrator. Sartain et al. (2001) reported that more observations or more observers resulted in more reliable evaluation scores. The finding from the study showed a .94 on the multi-facet Rasch analysis when looking at the average ratings from two evaluators, meaning there is high reliability. Both districts in the current study used the Danielson FFT.

Evaluation Practices for Michigan Teachers

In 2001, NCLB emphasized teacher quality, rather than teacher evaluation, as a part of certification and licensing. Michigan legislation recommended the use of quantitative data to improve student growth. The state required districts to ensure that a

substantial part of the evaluation rating was based on student growth. The Revised School Code Act 451 of 1976 established that performance evaluations for teachers include student growth with relevant data.

The MDE determined sound practices for conducting teacher evaluations. At the beginning of the academic year, the expectations of the evaluation system were communicated to teachers, allowing them to gain an understanding of the evaluation system. The intent was to minimize teacher anxiety by communicating the expectations. Evaluators would be more likely to ensure consistent practice and transparency by communicating the same expectations to all teachers. Schools and districts were expected to communicate the purpose for which the evaluation tool was to be used. If an evaluation tool was designed to assist teachers' professional growth but was being used solely for accountability purposes, this would leave out useful information for the teacher's development (Marzano, 2012).

The MDE (2011) recommended that principals designate a staff member to manage the principals' duties during times when observations were being conducted. This process allowed the principal to focus on the observation task, demonstrated to the teacher that the administrator was committed to completing a thorough observation, and helped to build trust between the teacher and administrator. To provide the best feedback, MDE recommended that conferences occur shortly after the classroom observation.

Models of teacher evaluation used in Michigan. Across the United States, different guidelines were established for districts to use in the selection of evaluation tools. Whereas some states had a statewide teacher evaluation tool, the Michigan

Department of Education did not dictate a statewide model. In 2011-2012, the Michigan legislature provided districts the autonomy to develop their evaluation system or to use evaluation systems already in existence as long as the established criteria were included. In 2012-2013, the Danielson FFT (Danielson, 2013), utilized by 488 Michigan public and charter schools, was the most widely used evaluation tool. The Marzano Evaluation Model (Marzano, 2012) was utilized by 176 schools in 2012-2013. A total of 346 schools utilized "other" frameworks that may have combined components from several different evaluation models.

In 2013-2014 Michigan adopted statewide evaluation guidelines as opposed to a specific evaluation model, as is required in other states. The Michigan Council for Educational Effectiveness established criteria that needed to be included in any district adopted evaluation system (MCEE, 2013). According to the Michigan Department of Education, Danielson FFT (2013), The 5 Dimensions of Teaching and Learning (Center for Educational Leadership, 2014), Marzano's Teacher Evaluation Model (Marzano, 2012), The Stages Online Evaluation Platform (Zimco, 2012), and The Thoughtful Classroom (Silver, Strong, & Associates, 2014) were the most widely used evaluation tools in Michigan (MDE, 2014). However, more than 200 districts in Michigan used internally developed models. Figure 2 depicts the teacher evaluation models and the number of schools that used each type of model.

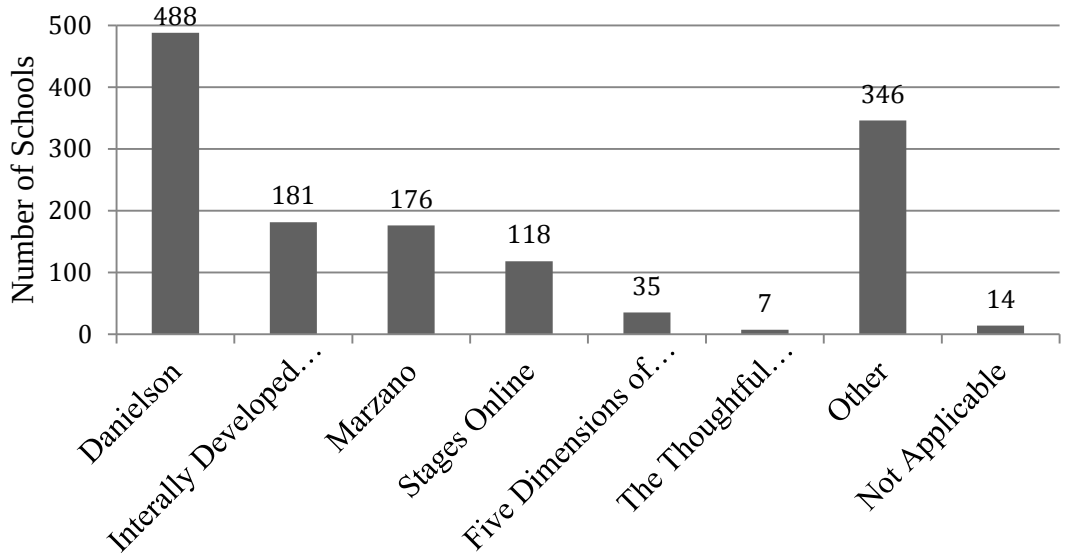


Figure 2: Observation Tools and Frameworks used to Evaluate Instructional practice (Michigan Department of Education, 2014)

The framework for teaching. Danielson's (2001) involvement in teacher evaluation provided insight into public education's goal to improve teacher quality continuously through the use of a comprehensive evaluation tool. The reform movements in use at the time of the current study were related to expectations in instruction and quality schools that were recommended in the 1983 publication, A Nation at Risk (National Commission on Excellence in Education, 1983). New Jersey, Illinois, Arkansas, Delaware, Idaho, South Dakota, Florida, and Washington adopted the Danielson FFT statewide for all school districts, with the New York City Public Schools, Los Angeles School District, Pittsburgh Public Schools also adopting this evaluation model (Danielson, 2013). This standards-based model used a 4-point rubric (Unsatisfactory, Basic, Proficient, and Distinguished) to evaluate each of the 22 components within the four domains of teaching responsibilities: (a) Planning and Preparation, (b) Classroom Environment, (c) Instruction, and (d) Professional

Responsibilities. The components aligned with those of the National Board of Professional Teaching Standards, which is responsible for nationally certifying teachers (Danielson, 1996). The evaluator collected evidence that served as documentation for the components, including observable student behaviors, work samples, and lesson plans on four domains. Table 4 illustrates the 22 components of the Danielson Framework for Teaching.

Table 4

Components of Danielson Framework for Teaching, Danielson (2013)

Domain 1: Planning and Preparation	**Domain 2: Classroom Environment**
• Demonstrating knowledge of content and pedagogy • Demonstrating knowledge of students • Setting instructional outcomes • Demonstrating knowledge of resources • Designing coherent instruction • Designing student assessment	• Creating an environment of respect and rapport • Establishing a culture for learning • Managing classroom procedures • Managing student behavior • Organizing physical space
Domain 3: Instruction	**Domain 4: Professional Responsibilities**
• Communicating with students • Using questions and discussion techniques • Engaging students in learning • Using assessment in instruction • Demonstrating flexibility and responsiveness	• Reflecting on teaching • Maintaining accurate records • Communicating with families • Participating in the professional community • Growing and developing professionally • Showing professionalism

Note. Adapted from Danielson Framework for Teaching, Danielson (2013), Available at www.danielsongroup.org.

Danielson (2013) identified research-based standards for quality teaching and established the importance of ensuring active student engagement to support high levels of learning. Teachers were provided with the Danielson FFT, which included an explicit guide that explained successful instructional practices. During an observation, evaluators expected to see teachers implementing the instructional practices outlined in Danielson's FFT. They used the Danielson rubric as the evaluation tool to assess teachers'

instructional effectiveness. For example, to receive the highest level on the rubric for the classroom environment, teachers needed to maintain classroom environments that were culturally sensitive and developmentally appropriate as well as provide opportunities for student ownership and make accommodations for individual student needs.

According to Danielson (2013), highly effective teachers were studied to understand how they interacted with students and colleagues, delivered instruction, prepared lessons, and enhanced their professional growth. In the field of evaluation models, the Danielson FFT is one of the most widely used tools. On its website the Michigan Department of Education (2013) listed the Danielson FFT as a resource to support school districts in their evaluation.

The 5 dimensions of teaching and learning. The 5 Dimensions of Teaching and Learning Evaluation Model was created by the faculty of the College of Education at the University of Washington (Center for Educational Leadership [CEL], 2014), which started out as a lesson analysis rubric. The descriptors provided evaluators with guidelines to look for in lessons. After extensive research on high-quality instruction, the tool evolved into an inquiry-based evaluation framework. At the time of the current study, the 5 Dimensions of Teaching and Learning Model (2014) was used to guide leaders in addressing the achievement gap by strengthening instructional approaches and emphasizing continuous improvement.

The CEL (2014) framework used a rubric with instructional descriptors and common language to create a vision for quality instruction in kindergarten through grade 12. The rubric was based on the student growth model and honored change in academic achievement over time (CEL, 2014). Teachers may have started the year low in one area

of the rubric, however, if they improved, the first score would not impact the final evaluation score. Self-assessment was used as an ongoing practice within the instructional framework. Through self-reflection instructional focus areas were identified and goals were set. Teachers' goals were supported through professional development opportunities. Table 5 illustrates the five components of the 5 Dimensions of Teaching and Learning (2014).

Table 5

5 Dimensions of Teaching and Learning

Purpose	Student Engagement	Curriculum and Pedagogy	Assessment for Student Learning	Classroom Environment and Culture
Planning with alignment to standards, setting clear expectations	Encouragement through challenge and intellectual thinking	Ensuring that instruction challenges and supports all students	Using ongoing assessment to shape and individualize instruction	Creating community, equity, and that maximize opportunities for learning and engagement.

Note. Center for Educational Leadership, 2001, available from https://www.k-12leadership.org/ content/service/5-dimensions-of-teaching-and-learning

Schools and districts that used the 5 Dimensions of Teaching and Learning (2014) scored the criterion on a 4-point rubric, from unsatisfactory to distinguished. With the growth model related to student achievement, an additional component was added. The growth score was calculated with the other five areas for the final summative rating (Fink, McDermott, Austin, & Cloninger, 2001)

Marzano teacher evaluation model. Starting in 2003 with the Study of School Effectiveness, Marzano's research continued through 2005 with a meta-analysis of school leadership. The development of the Marzano Evaluation Model was an outcome of the 2010 study of *What Works in Schools* (2010). The Marzano model used a 5-point scale

to rank the components under each domain. The range started with a score of zero, meaning strategies were not evident, and increased, with descriptors, to four points, the highest points possible (Marzano, 2012). The model's four domains included the overarching non-negotiable goal to increase student achievement. The four domains Marzano lists in the Marzano Evaluation Model include Behavior and Strategies, Planning and Preparation, Reflecting on Teaching, and Collegiality and Professionalism. Design questions were used in this model to help teachers think through their instructional practices to intentionally enhance learning. Though the Marzano model does not have a separate student growth category the evidence from the other domains can be correlated with student growth.

Domain 1. - behaviors & strategies. The classroom behaviors and strategies domain refers to instructional practices teachers use in the classroom to enhance learning. There are 41 strategies and actions within domain one. Tracking student progress, celebrating success, providing clear learning targets and classroom routines are a few of the components expected to be implemented on a regular basis. Some examples of content related components are learning reflections, practice strategies for deepening understanding, and organization for cognitively complex tasks (Marzano, 2012).

Domain 2. - planning and preparation. The planning and preparation domain requires teachers to plan thoughtfully and prepare lessons to meet the required standards and objectives. Teachers need to consider the differentiated needs of all students, from the needs of English Language Learners to special education students, to gifted learners, to those students who lack support from home and are at-risk of failing. The Marzano model promotes scaffolding instruction and making effective instructional decisions with

learning gains as the result (Marzano, 2012).

Domain 3. - reflecting on teaching. The reflecting on teaching domain requires teachers to be self-aware and make action plans for instruction, leading to continual growth. Teachers are expected to analyze student work to identify pedagogical strengths and weaknesses. The outcome of Domain 3 connects professional development and collaboration (Marzano, 2012).

Domain 4. - collegiality and professionalism. The collegiality and professionalism domain focuses on school leadership and development of individual teachers through collaboration. The objective is for continuous improvement to become the culture within the building. The environment that was created through promoting a positive culture enhances and impacts classroom strategies and behaviors. Teachers collect artifacts to show how they participated with colleagues by exchanging ideas and helping each other to attain their goals (Marzano, 2012).

The thoughtful classroom. Silver, Strong, and Associates (2014) developed The Thoughtful Classroom Teacher Effectiveness Framework. This model was the result of a study based on 35 years of research in over 2,500 schools. The framework, with three components (Effective Classrooms, Instructional Design and Delivery, and Professional Practice: Looking Beyond the Classroom) created a standard language about quality teaching (Silver, Strong, & Associates, 2014). Classroom instruction was the focus of two of the three components. As defined in this framework, Thoughtful Classrooms are organized, encourage positive relationships, and engage students in a culture of thinking. Preparation for learning, reinforcing deep understanding, application of knowledge, and reflecting and celebrating learning were subsets of the second component of this

framework, Instructional Design, and Delivery. The third part was Professional Practice: Looking Beyond the Classroom, this component focused on the teachers' commitments to growth and continuous learning, contributions to the school community, and professionalism (Silver, Strong, & Associates, 2014).

Principals observing teachers using this model focused on asking essential questions after the lesson that would allow the teacher to reflect on the lesson components. Principals may question teachers to explain how useful they think they were in activating students' prior knowledge. A principal might ask the teacher's thoughts on other ways that could have prepared the students for the lesson. Rather than simply scripting a lesson, the principal records evidence during the observation, which supports the essential questions. The feedback that is collected and shared with the teacher follows a structure called the Four Ps: provide evidence to support what was observed, praise for positive impacts on student learning, pose questions for reflection, and propose ideas to improve the teacher's practice. A 4-point rubric distinguished teachers as a novice, developing, proficient, or expert based on the overall points. The rubric was used for each descriptor within a component level as well as for the overall ranking (Silver, Strong, & Associates, 2014).

STAGES evaluation tool. A Supportive Tool for Assessing Growth in Educational Systems STAGES (Zimco, 2012) was developed by Saginaw Valley State University in collaboration with several Michigan School Districts in the Saginaw area. Though listed on the Michigan Department of Education as a teacher evaluation model, STAGES is a data warehouse used to maintain, store, and track progress. This program allows each district to customize their tool to match their district's specific needs.

Districts that have their own evaluation tool may use STAGES as the vehicle to transform their tool into a web-based model. At the time of the current study, 118 districts in Michigan were using STAGES (Zimco, 2012).

Districts' evaluation processes. In the present study, a review of teacher evaluation systems was conducted in two different school districts in Michigan. Both schools' evaluation models were revised to meet Michigan's legislative requirements for the evaluation process (Act 451, 380.1249, 2009). District B reviewed their evaluation process in 2012 and District A in 2013. After committee meetings comprised of teachers and administrators, both districts chose to adopt the Danielson FFT as the district evaluation tool. Danielson's FFT Evaluation Model produces reliable scores. This statement is based on reliability studies by Sartain, Stoelinga, Brown (2011) and Milanowski (2011). Highly effective, effective, minimally effective, and ineffective were the four rating levels used by evaluators to rate a teacher's end-of-the-year performance. Teacher effectiveness ratings were determined through a series of classroom observations; both formal, or scheduled, and informal, unscheduled; student performance on local and state assessments; and self-reported information such as lesson plans, survey results, and other evidence of effectiveness levels. The specific instructional categories measured in both districts were: instruction, planning and preparation, classroom environment, professional responsibilities, and student growth. Within each category or domain, some components were evaluated and scored. Teachers could earn a 4, 3, 2, or 1 on the components under each domain. Each domain was given a separate score made up from the component scores for the corresponding domain. In both districts, the domain scores were weighted. As mandated by the state, student growth was to have a higher

weight than any of the other domains. The total of the weighted domains generated a cut score. The effectiveness ratings for both of the districts studied were assigned according to the following cut scores: 4.0-3.5 highly effective, 3.49-2.75 effective, 2.74-2.0 minimally effective, and 1.99-0.0 ineffective. These cut scores were locally established by the districts' evaluation study teams (personal communication, July, 2014).

A holistic definition for each of the four teacher evaluation rating levels does not exist for the state of Michigan. The largest gap exists between the number of teachers rated highly effective and teachers rated effective (Keesler & Howe, 2012). Educators who review the data should be aware that the sample for minimally effective and ineffective rated teachers is small for districts in the study and across the state. The majority of teachers across both districts and the state of Michigan were rated either highly effective or effective for their overall scores (MDE, 2013).

Building level teacher evaluation process. Goal setting is an important part of the evaluation process (Tucker & Stronge, 2005). At the beginning of the year, the teacher and administrator meet to review goals, ensure early implementation and make revisions, if necessary (Tucker & Stronge, 2005). Administrators are responsible for ensuring that measurable goals are developed as part of the evaluation process. The follow-through resulting from this meeting should hold all stakeholders accountable for goal attainment. Some evaluation models require teachers to rate themselves on the evaluation model criteria and establish goals for themselves. Through self-reflection, teachers can establish professional development goals for the school year.

Most teacher evaluation systems rely on a limited number of principal observations throughout the school year. The traditional approach to the observation

process usually begins with a pre-observation meeting between the teacher and the evaluator. The teacher might explain the lesson plan and objectives, clarify what took place the previous day, or may discuss individual student needs. This conversation between the evaluator and the teacher is followed by the classroom observation. During the observation of the lesson, the evaluator scripts what took place. The evaluator is looking for particular instructional, environmental, and procedural strategies that the teacher demonstrates during the lesson. The evaluator also observes student and teacher relationships, as well as student-to-student interactions (Danielson, 2013).

The teacher-principal debriefing session allows the teacher to reflect about the lesson, objectives, and any adjustments that were made during the lesson based on student responses. The quality of the debriefing session depended on the level of questioning posed by the evaluator. Low-level or basic questions resulted in less meaningful conversations (Sartain et al., 2011). It is appropriate for student work to be shared during this meeting as evidence of student understanding. The researchers in the Chicago School Research Study (Sartain et al., 2011) gathered data on the process of teacher evaluation using teacher surveys to collect input from teachers on the value of the post-conference process. The debriefing session should not be completely dominated by administrative feedback, rather the primary focus should be teacher reflection (Sartain et al., 2011). Marzano (2012) believed that teacher development and a focus on instructional growth should be the priority of evaluation systems. If accountability and measurement are the focus of evaluation, the principal shares the scores or points the teacher earned on lesson components. Recommendations may be made for professional development opportunities, such as observing a colleague focus on a particular area

within the criteria on the rubric.

This same process is followed for additional observations throughout the year. Though there is not a formal board policy in either of the districts used for this research, districts require principals to complete scheduled and unscheduled observations on each teacher on a yearly basis. At a minimum, districts must follow state guidelines for the number of observations. Michigan legislation requires all teachers, unless they earn a rating of effective or highly effective in their two most recent annual year-end evaluations, to be observed by the evaluator, at least, two times each school year (Michigan Revised School Code of 1976, 2014). Districts may choose to exceed the state expectations. At the end of the year, a final evaluation score is calculated based on the preponderance of evidence collected throughout the year.

According to Danielson and McGreal (2000), teachers and administrators comply with the district mandates of the evaluation process, but some teachers feel they do not gain much from the experience. Some teachers feel that their evaluator is not knowledgeable in their particular content area, and, therefore, they might not respect the administrator's feedback. Implications of this process make teacher evaluations useless therefore having no impact on student achievement.

Increased Accountability

After A Nation at Risk (National Center for Education Evaluation [NCEE], 1983) had been presented to the public, the focus of education turned from merely presenting content to students to holding educators accountable for student learning. Two important outcomes resulted in the release of A Nation at Risk. The first outcome was the NCEE demand for more rigorous educational standards, leading to a focus on standards-based

educational reform. Title I funds were used to support the demands outlined in A Nation at Risk. In connecting with teaching and learning, the second outcome focused on revising teacher evaluation systems. Public school institutions needed to be held accountable for the public funds they used, and more importantly, the futures of the students they taught (Supovitz & Poglinco, 2001).

In 1994, the reauthorization of the Elementary and Secondary Education Act (ESEA) of 1965 emerged as the Improvement of America's Schools Act (IASA) under President Clinton's administration (Riley, 1995). Continuing to require all states to create rigorous standards for reading and writing, the IASA also required all states to administer statewide assessments. Assessments were required to be conducted at least one time per year at all levels. Waiver provisions were permitted as a result of IASA. School districts could request waivers from the United States Department of Education if they found more effective ways to meet requirements (ESEA, 1994).

Accountability and standards-based reform would take a few more years before becoming the focus of the country. President George W. Bush, in 2001 enacted the No Child Left Behind Act (NCLB) that would influence another decade of reform. Teacher evaluation systems were impacted by the NCLB mandate.

> With the adoption and implementation of No Child Left Behind (U.S. Department of Education, 2002) and the resultant call by the National Governors Association (NGA) to target teacher evaluation policy as a way to achieve the goal of a highly qualified teacher in every classroom, policy makers focused efforts on this promise to improve student learning. (Goldrick, 2002, p. 2).

NCLB established an expected nationwide target for the year 2014. The AYP provision of the law required students to make individual gains each year in the areas of mathematics and English language arts, with all students scoring proficient in both areas by the year 2014. The target was established indicating that all students would be proficient or exceed proficiency as measured by state assessments. Administrators would now be charged with holding teachers accountable to these high expectations. Teaching based on student achievement became the primary emphasis. Teachers would be required to establish specific, achievable goals for students to improve performance (Toch & Rothman, 2008).

With state test scores in the spotlight, a shift in teacher evaluation tools was necessary. Previous teacher evaluation instruments failed to link student achievement to teacher evaluations. The new tools needed to explain quality teaching criteria clearly and focus on student growth. Though guidelines were established, NCLB did not mandate a particular tool to determine teacher effectiveness. School districts or states had the autonomy to create evaluation instruments that met the needs of their local districts (USDE, 2002).

Equally important, the NCLB legislation introduced the requirements necessary to be considered a highly qualified teacher (USDE, 2004). The highly qualified provision required teachers to be considered an expert in the field they taught. At the time of the current study, to be considered highly qualified, teachers were required to hold a bachelor's degree and a state certification or license for the grade level or subject matter taught. Also, highly qualified teachers were required to demonstrate their subject-matter and content expertise. Experienced teachers reported this highly qualified status through

administrator signed professional development and experience logs, passing scores on state-developed tests, graduate degree status, or through a state-approved content expertise process (USDE, 2004). Experienced teachers mentored teachers with fewer years of experience. Mentoring and professional development logs were kept to meet the highly qualified criteria (USDE, 2004). In addition to teacher certification requirements, NCLB legislation required states to establish quality teacher guidelines. "Public education defines teacher quality language in terms of credentials teachers have earned, rather than on the basis of the quality of work they do in classrooms" (Toch & Rothman, 2008 p. 2). Districts were held accountable for monitoring the guidelines for high-quality teachers. District administrators moved quickly to create methods to measure teacher performance in association with student performance on state reading and mathematics assessments. The link between student scores and teacher performance would be irrevocably linked (Braun, 2005).

In 2009, Race to the Top was introduced to support NCLB. An outcome of the American Recovery and Reinvestment Act (ARRA), Race to the Top was an education grant challenging schools to use student data as part of the overall teacher evaluation rating system. Although research is limited in supporting teacher performance based on student growth data, additional states were requiring these data as a component of teacher evaluation. Studies determined that teachers who earned higher evaluation scores had students achieving higher academic scores (DiPaola & Hoy, 2014; Marzano, 2011). Teacher effectiveness based on student performance was implemented in many school evaluation policies as an expectation for a portion of the teacher's aggregate evaluation score.

Accountability through supervision. The role of accountability increased with the enactment in 2001 of NCLB. The emphasis from NCLB was on teacher quality as a key factor in improving student achievement. With this came an increased emphasis on ensuring that highly qualified teachers were in every classroom. While Michigan is one of 17 states where teachers were being held accountable in their overall evaluation of their students' performance levels, limited studies were published that supported effectiveness of teacher evaluation models using student performance as a major indicator for identifying a highly effective teacher. Schools were measuring progress on locally created common and formative assessments. Though inconsistently measured from district to district, the student growth component of the evaluation tools in Michigan were required to carry the heaviest weight of all evaluated components (Keesler & Howe, 2012).

Equipping teacher leaders and administrators with guidelines and strategies for conducting effective classroom visits were ways to improve practice. Frequent, short, unannounced classroom visits were conducted to determine if teachers were selecting rigorous questions, communicating learning targets, having students self-assess their work, and providing students with actionable feedback (Behrstock-Sherratt et al., 2013). The use of scheduled, less frequent observations were likely not as authentic as parallel forms of observation. Knowledge of a scheduled evaluation often resulted in a showy, somewhat staged lesson. These pre-planned lessons tended to look different from unscheduled visits by the evaluator (Marshall, 2012).

Studies of Teacher Performance and Student Achievement

Empirical research has been conducted on the relationship between teacher

performance and student achievement. Mooney (2012) found in the "typical district: 20% of teachers are 'highly effective,' 60% to 65% are 'effective,' 10% are 'partially effective' and 3% to 5% are 'ineffective'" (para. 26). Summarized in this section are multiple studies regarding the relationship between teacher performance and student achievement.

Heneman, Milanowski, Kimball, and Odden (2006) focused on the relationship between the teacher evaluation scores using the Danielson standards-based teacher evaluation system and student achievement scores. Heneman et al. (2006) sought to determine if teachers and administrators believed the system for evaluation was fair and accurately guided teachers' efforts to improve instruction. Heneman identified multiple years of data were analyzed in four Cincinnati schools. Student scores over the 3-year study showed positive increases in the schools, where grade levels ranged from second to eighth grade. All four schools showed a variety of increases in test scores, which were attributed to the differences in evaluators, socio-economic status, training, the number of students, and ethnicity. Table 6 depicts the correlations between teacher evaluation ratings and student achievement in reading and mathematics. The results from the Heneman study revealed the standards-based evaluation systems when used appropriately, could have a positive influence on student achievement when instructional strategies that are measured by the evaluation tool also measure student learning.

Table 6

Average Correlations Between Teacher Evaluation Ratings and Student Achievement in Reading and Mathematics

School	Year	Grades	Reading	Mathematics
Cincinnati				
	2001-2002	3-8	.48	.41
	2002-2003	3-8	.28	.34
	2003-2004	3-8	.29	.22
	3-year average:		.35	.32
Coventry				
	1999-2000	2,3,6	.17	.01
	2000-2001	2,3,4,6	.24	-.20
	2001-2002	4	.29	.51
	3 year average:		.23	.11
Vaughn				
	2000-2001	2-5	.48	.20
	2001-2002	2-5	.58	.42
	2002-2003	2-5	.05	.17
	3-year average:		.37	.26
Washoe				
	2001-2002	3-5	.21	.19
	2002-2003	4-6	.25	.24
	2003-2004	3-6	.19	.21
	3-year average:		.22	.21

Note. Adapted from *Standards-Based Teacher Evaluation as a Foundation for Knowledge- and Skill-Based Pay* (Heneman, Milanowski, Kimball, & Odden, 2006)

Allemann (2006) investigated a California Elementary School that sought to determine the connection between the teacher evaluation process and increased student achievement. School data for California Elementary School in California City, California indicated 95% of the teachers were "highly qualified" under NCLB. The majority of the students (89%) were considered economically disadvantaged and achievement scores either met or exceeded the growth targets. Based on the student achievement scores and the high number of economically disadvantaged students, the school was outperforming its demographics. In addition to gathering input from a teacher survey, data were

collected through observations in the school and interviews with the teaching staff. Through the observations, it became evident that collaboration around student data and interventions were valuable practices within this school. Teachers engaged in professional dialogue regarding teaching and learning. Allemann observed, during classroom visits, the application of instructional strategies that emerged from the collaborative meetings.

Professional development surfaced as a significant focus for the teachers at California Elementary School (Allemann, 2006). Teachers had a clear understanding of areas where they needed to improve. Additionally, the methods to improve through professional development were differentiated, meaning every teacher had different needs and could attend professional development based on their needs rather than having to participate in professional learning that was not essential for their growth. The teachers and administrators worked together to seek alternative ways to support individual teacher growth.

According to Allemann (2006), the results of the study also revealed a central theme of shared leadership. The principal acted in partnership with the teachers. The leadership was non-threatening with the principal facilitating conversations promoting reflective practice and collaboration among teachers. The teachers were important contributors to the direction in the building. The study also revealed that high expectations were set for all students at the California Elementary School. Teachers communicated the expectations to students and teachers were expected to support students as they worked to reach the academic goals. Findings from Allemann's study, particularly teacher surveys, revealed that the evaluation process had an indirect

association with improved instruction. The process of collaboration, supportive leadership, differentiated professional development, and high expectations had the greatest influence on student achievement improvement.

Rockoff and Speroni (2011) studied the effects that new teachers had on student achievement in a quantitative study using New York City teachers. In the study, teacher performance was measured through student assessment results and observational data. Student growth was measured through multiple assessments including national, state, and local assessments, as well as student work samples. Teachers were involved in determining the criteria that were used to monitor student growth. The researchers concluded that subjective evaluations and objective performance data were essential for a quality educator evaluation system. Evaluation data from over 4,000 mathematics and English teachers were used in the study. The data were sorted into three areas: (a) subjective evaluations performed by the new teacher's mentor, (b) subjective evaluations from certified evaluators, and (c) objective evaluations determined by student achievement scores from 2003-2008. The results of the study revealed that teachers who received higher subjective evaluations in their first year of teaching or participated in a mentoring program produced higher gains in student achievement. First-year teachers with higher student achievement gain tended to produce even greater gains in year two.

In each of the previous studies, various indicators linked student achievement data and teacher evaluation ratings. In the state of Michigan, multiple forms of data were used to determine student growth and how it related to teacher evaluation ratings. Table 7 shows the percentages of student growth measures used in Michigan for educator evaluations in kindergarten through grade 8 from 2011-2012 and 2012-2013.

Table 7

2011-2012 and 2012-2013 K-8th Grade Student Growth Measures Used in Educator Evaluations K-8

Assessment	Percentage of Districts Using Assessment 2011-2012	Percentage of Districts Using Assessment in 2012-2013
State Assessment	72.9%	61.9%
Local Assessment	68.2%	60.8%
Dynamic Indicators of Basic Early Literacy Skills (DIEBELS)	57.3%	52.1%
Student Work Sampling	39.3%	37.4%
American College Testing (ACT) Explore	38.5%	33.5%
Northwest Evaluation Association (NWEA)	23.9%	33.0%

Note. Adapted from *Michigan Educator Evaluations & Effectiveness Report* by *(MDE)*, 2013-2014, Retrieved from https://www.michigan.gov/documents/mde/2013-14_Educator_Evaluations_and_Effectiveness_485909_7.pdf

Improving Teacher Evaluations

Many authors have suggested that teacher evaluation systems are flawed. Hull (2011) wrote, "There is a huge variability among teachers, even within schools, but it is hidden by inadequate evaluation tools" (para. 2). Minimally effective teachers can have data that shows overall class growth and achievement, the opposite is also true (Shavelson et al., 2010). There are multiple reasons why weak teacher evaluation systems cause evaluators to struggle. Not all districts use the same assessment tool and although some districts use the same evaluation tool; the way the tool is used may differ. Various districts only use the bottom three performance levels; this eliminates any teacher from attaining the highest performance label. Goal setting for teachers can also vary depending on the content and grade level taught. Due to the inconsistencies teachers

may be rated highly effective in one district; however, if that teacher moves to another district with different criteria for the teacher rating levels, the teacher may not have the same effectiveness rating (personal, 2014). In 2009, the Center for Education found that an overwhelming majority of teachers met the law's definition of highly qualified, yet there was little indication that teacher quality had noticeably improved (Center for Education, 2009). These are just a few examples of the issues with teacher evaluation systems.

Improving teacher evaluations starts with setting clear expectations that encompass all factors of teaching. These expectations should be communicated through a systematic evaluation tool. Milanowski, Kimball, and White (2004) indicated teacher practices that happened outside of the classroom, such as planning, professional development, collegial contributions, parent communication, and evidence of social and emotional growth for students should be considered when developing evaluation tools. Based on changes by policymakers, educators were spending more time reevaluating and redesigning their teacher evaluation systems to strengthen instructional practices. As part of this process, school districts were trying to find ways to support alignment of instructional development with increased demands for accountability. Completing the evaluation cycle or looking at a single test score was not enough to determine teacher effectiveness (National Education Association [NEA], 2010).

Using assessments that measure teachers' influence on student achievement can determine teacher competence and direct the focus to areas where teachers need additional support. Concerns have been raised about using assessment scores to measure teacher performance. Factors that cannot be controlled such as a student's home

environment, the influence of other teachers on achievement, and individual student characteristics all contribute to the one final test score (Darling-Hammond, 2010).

Summary

Accountability measures for student growth established through political policy have increased in school districts across the United States. Through the review of literature, the research on accountability in education from the evolution of testing to teacher evaluation systems. Studies revealed that highly qualified teachers had a positive impact on assessment scores over time, regardless of students' socioeconomic status.

The purpose of the literature review was to provide the reader with an understanding of the history of standardized testing and the connection between public school accountability and testing. Additional information was provided regarding the teacher evaluation tools used in schools at the time of this study. Studies on student achievement and teacher performance ratings were provided to help illustrate the legislative requirements for incorporating student achievement performance into the teacher evaluation ratings. The content of chapter three includes methodology used in the study as well as the population and sampling procedures, instrumentation, measurement, validity and reliability, data collection conducted to determine the relationship between teacher evaluation scores and student achievement, data analysis, hypothesis testing, and limitations.

Chapter Three

Methods

The purpose of this study was to determine if highly effective and effective teacher evaluation ratings and teacher experience have an impact on student achievement for third and sixth-grade students as measured by the fall Michigan Educational Assessment Program (MEAP) in mathematics and English language arts. Described in this chapter are the methods that were used to collect and analyze the data. The topics outlined include research design, population and sample, sampling procedures, instrumentation, data collection procedures, data analysis and hypothesis tests, and limitations.

Research Design

A non-experimental, ex-post facto research design was used for this comparative data study (Lunenburg & Irby, 2008; Johnson & Christensen, 2008). In this study, the independent variables were not manipulated, and the researcher did not randomly assign participants to the study groups. Therefore, the design of the study was non-experimental (Johnson & Christensen, 2008). The data for the study were obtained from past school records, with no data collected directly from the teachers in the two groups. The study groups used in this research were second and fifth-grade teachers who were rated as either effective or highly effective on their annual spring evaluation for 2013 and 2014. The teachers were further categorized by their years of experience, new teachers, level I experienced teachers, and level II experienced teachers. The independent variables were teacher performance and teacher experience. The dependent variables were the classroom scaled scores for the third and sixth-grade fall MEAP mathematics and English

language arts scores. Student data from fall MEAP scores were linked to the teacher who taught the students the previous year, when the students were in second and fifth-grade. The teacher evaluation data came from the second and fifth grade teachers, the MEAP data came from third and sixth-grade student scores. The data was drawn from two districts in suburban Detroit.

Population and Sample

A target population is utilized when it is not possible to gather data from the larger group (Creswell, 2014). The target population of this study was Michigan elementary teachers who were rated highly effective and effective on their annual evaluations. The focus of the study was on a subset of elementary teachers from two suburban Detroit school districts. A breakdown of the number of teachers whose data was used is shown in Table 8.

Table 8

Sample of Teachers Included in the Study

Academic Year	2nd Grade Teachers	5th Grade Teachers
2012-2013	45	40
2013-2014	47	49

The study group was selected using a purposive sample that included second and fifth-grade teachers from 15 elementary schools in two suburban Detroit school districts. To determine if conclusions could be made between primary; kindergarten, first, and second-grade teachers, and intermediate; third, fourth, and fifth-grade teachers, the researcher chose to look at 2012 and 2013 data from second and fifth-grade teachers. The sample included data from the same teachers for both years unless a teacher left the

school, changed grade levels, or received an evaluation lower than an effective rating. Kindergarten and first-grade teachers were not considered for this sample since students do not take the MEAP until third-grade. Teachers who were rated minimally effective or ineffective were not part of this study.

Sampling Procedures

Data were selected through purposive sampling. "Purposive sampling involves selecting a sample based on the researcher's experience or knowledge of the group to be sampled" (Johnson & Christensen, 2008, p. 175). The current study included 85 second and fifth-grade teachers in the 2012-2013 sample, the 2013-2014 sample included 96 second and fifth-grade teachers. Teachers who were rated as either highly effective or effective on their annual evaluations were included. Archived data for these teachers were obtained from both school districts. Each district's assessment department provided archival data for these teachers.

Instrumentation

According to school officials the MEAP assessment linked previous grade level content to current grade level content. In mathematics, students answered multiple-choice questions. These questions were designed to measure mathematics content expectations. The assessment for grade three had 63 questions. A variety of questions given at each tested grade level came from the following five strands: Numbers and operations, algebra, geometry, measurement, and data and probability. The same strand names are used for each grade level. Using graphs, pictures, story problems, or data tables, students responded to multiple-choice questions, each worth one point. These concepts closely matched the standards from the National Council of Teachers of

Mathematics. Content knowledge and application of concepts were measured throughout all grade levels tested on the MEAP (MDE, 2010).

The MEAP English language arts test requires students to read for understanding across multiple texts, use text features, knowledge of genres, and text structures to construct meaning from themes within the text. Students respond to both multiple choice and constructed response questions. The MEAP multiple-choice items require students to select one correct response from the options provided. On the multiple-choice items, students receive one point for correct answers. The multiple-choice items are machine scored.

To prepare for the testing process, teachers can use released sample items, for all content areas tested in the MEAP. The MEAP is the only test that measures what Michigan students should know and be able to do against established Michigan content standards (MDE, Office of Assessment & Accountability, 2004-2005). The results of the MEAP test indicate the level of proficiency that a student demonstrated: Advanced (Level 1), Proficient (Level 2), Partially Proficient (Level 3), and Not Proficient (Level 4). The goal is for students to score proficient or above. As a criterion-referenced test, mastery of standards are based on grade level content expectations (GLCEs) and are monitored through MEAP results. Criterion-referenced tests (a) require test takers to answer the same questions, or a selection of questions from a standard set of questions, in the same way; and (b) are scored in the same manner, making it possible to compare the performance of individual students or groups of students. (MDE, 2011).

Constructed-response items are handwritten by students, and they are required to be hand scored by trained scorers. Before teachers are allowed to score this portion of

the assessment, they must pass a rigorous training. The scorers used a detailed scoring rubric that aligned with the state standards (MDE, 2011). A single score is given even though multiple criteria may be measured. Each grade level has a section that assesses students' knowledge of word recognition. The English language arts assessment measures performance on the Michigan Grade Level Expectations for Reading (MDE, 2011).

Measurement. The independent variables were teacher performance as measured by being rated as either highly effective or effective in their annual evaluations and teacher experience as determined by the number of years they had been in their school districts

- new - 0 to 3 year
- level I - 4 to 12 years
- level II -12 years of experience or beyond

The dependent variable was the fall MEAP scaled scores for the third and six-grade reading and mathematics tests that were averaged for each teacher in the study.

Validity and reliability. When a study has a reliable data collection process, it can be repeated. According to Lunenburg and Irby (2008), "Reliability is the degree to which an instrument consistently measures whatever it is measuring" (p. 182). Lunenburg and Irby (2008) stated, "Validity is the degree to which an instrument measures what it purports to measure" (p. 181).

Danielson Framework for Teaching (FFT). The Chicago Consortium of School Research (Sartain, L., Stoelinga, & Brown, 2011) found the FFT to be a valid measure of teaching practice. To determine validity, quantitative and qualitative research was conducted using the FFT. Principals observed 757 teachers in the Chicago Public

Schools. The study findings indicated that students who showed the least growth in test scores were instructed by teachers who earned the lowest ratings on the FFT (Sartain et al., 2011). The greatest gain in student test scores was observed in classrooms where teachers received the highest ratings on the Danielson FFT (Sartain et al., 2011). Figure 3 provides the percentage of growth by the teachers' evaluation ratings.

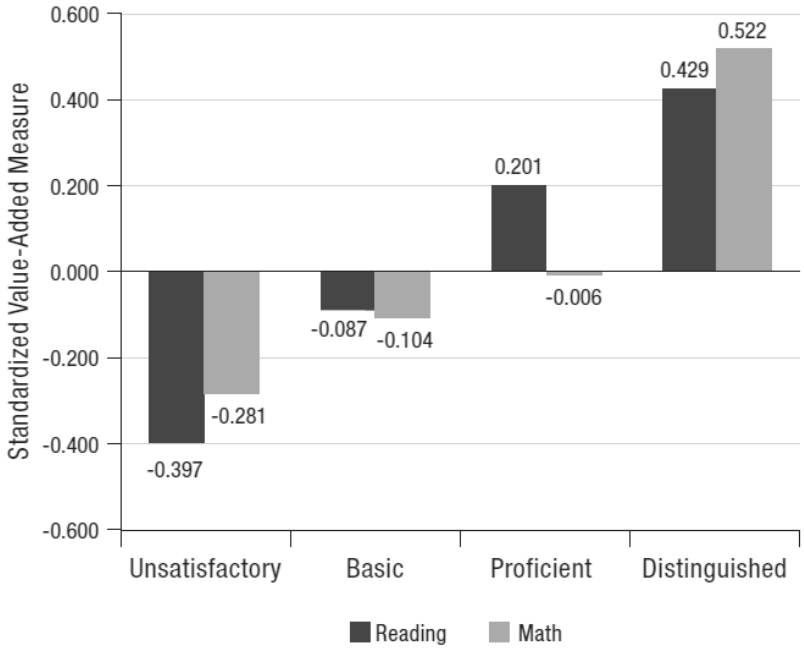


Figure 3: Percentage of growth by the teachers' evaluation ratings

Note. Sartain et.al., 2011 Reprinted from https://consortium.uchicago.edu/sites/default/files/publications/Teacher%20Eval%20Report%20FINAL.pdf

MEAP. Test validity has been an ongoing process that started when the MEAP test was initially developed and will continue until the test is no longer being used. Content and curricular validity for the MEAP assessment assures educators that the test content accurately assesses the state standards that were to be measured. From the beginning stages of MEAP development statewide, assessment teams consisting of item

development experts, assessment experts, and Bureau of Assessment and Accountability (BAA) staff worked collaboratively to study the tests. Annually, the team participated in reviewing field-tested and new items for the MEAP assessment (MDE, BAA, & Measurement Incorporated, 2011-2012).

The assessment team reviewed test items for difficulty, appropriateness, and fairness, in addition to checking for alignment to the standards that the items were intended to measure. By ensuring that the test items were relevant and aligned with the standard to be measured, the assessment team provided evidence to support the validity of the MEAP results. Test items that were not aligned with the content standards could be resubmitted after revisions were made. When items were approved, they were put into the MEAP test. Without approval, the alignment of test items to content objectives would not be valid measures. Having detailed review procedures provided confidence in the validity of the MEAP results (MDE, BAA, & Measurement Incorporated, 2011-2012). Figure 4 shows the MEAP Item Develop/Review Cycle.

MEAP Item Development/Review Cycle

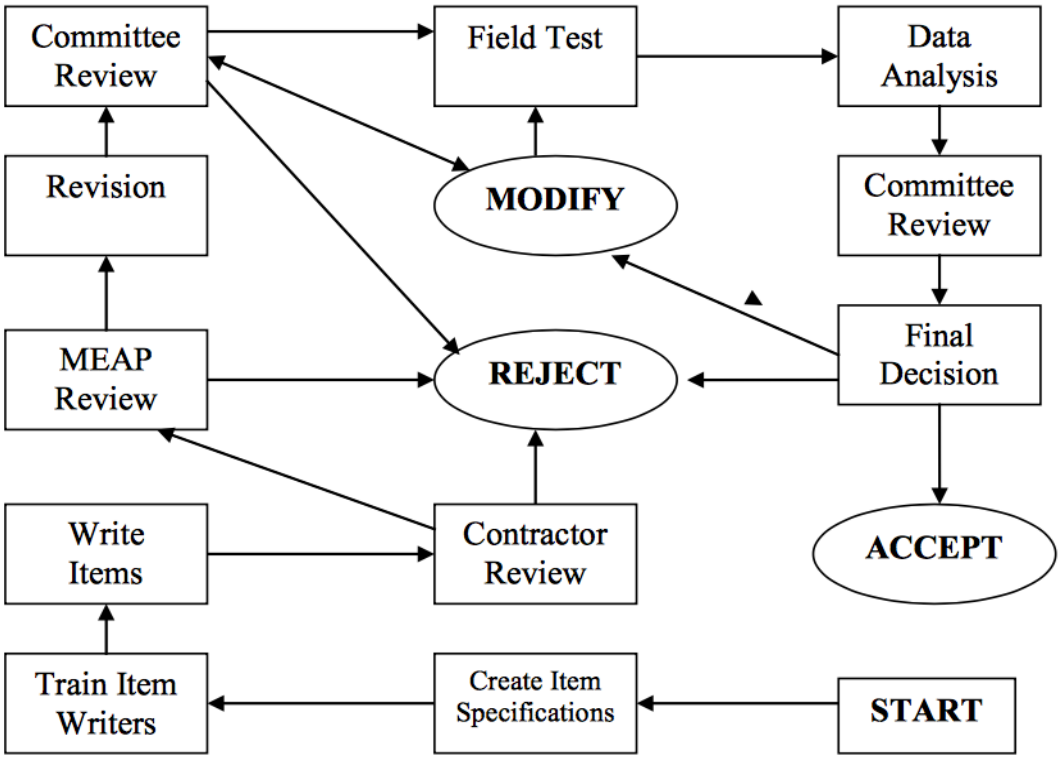


Figure 4. Shows the procedures that are followed to ensure content validity on MEAP assessment items. Adapted from Michigan Department of Education, Bureau of Assessment and Accountability & Measurement Incorporated (2011-2012). *MEAP Technical Report*. Lansing: Retrieved from http://www.michigan.gov/documents/mde/MEAP_2010-2011_Technical_Report_394693_7.pdf

MEAP alignment studies for English language arts and mathematics were conducted in 2005 in Lansing, Michigan (MDE, BAA, & Measurement Incorporated, 2011-2012). The English language arts reviewers met to analyze the state content standards and MEAP assessments for third through eighth grade. The team consisted of eight assessment experts from Michigan, and four assessment experts from other states. Each of the experts used the same criteria to review the test. The experts noted a few standards were not assessed on the test, whereas other standards were assessed multiple

times from one grade level to another. Although two objectives for reading were over-emphasized, the assessment team found the test to be valid and did not require any items to be rewritten (MDE, BAA, & Measurement Incorporated, 2011-2012).

The same process was completed for the mathematics MEAP test. The assessment team consisted of 13 reviewers. In addition to the content experts, district mathematics supervisors, mathematics teachers, and a mathematics professor participated in the review. Three experts were from states other than Michigan. The participants reviewed assessments for third through eighth grades. According to the assessment criteria review, conducted by this group of experts, all but one grade level was fully aligned with the standards. The third-grade assessment required modification to the data and probability content strand. A large number of items aligned only to one standard, with this discrepancy requiring six items to be replaced for the third grade MEAP assessment (MDE, BAA, & Measurement Incorporated, 2011-2012). At the time of this study, the MEAP assessment did not have any criterion-related validity reported. The technical report from 2011-2012 stated that evidence was collected and reported on an annual basis. However, MDE did not publish any updated technical reports beyond the 2011-2012 version.

Reliability. Reliable data provide teachers with direction and focus. When the data are reliable, educators can use the numbers to determine what type of instruction is needed and specifically in what areas. The Danielson FFT model was used to evaluate teachers' performance. The Danielson FFT is an open-ended evaluation instrument that administrators use to rate the teachers' performance in four domains, including planning and preparation, classroom environment, instruction, professional responsibility. The

Danielson FFT model was tested for inter-rater reliability (Milanowski, 2011). The purpose for using inter-rater reliability was to provide assurances that the Danielson FFT was measuring teacher performance. Two raters were used to determine the inter-rater reliability of observations of 99 teachers in Cincinnati. Milanowski (2011) reported that inter-rater reliability was low on Domain 2 - Classroom Environment, (73%) and Domain 3 - Instruction, (79%). The results of the testing provided support that inter-rater reliability was low, but Milanowski explained possible reasons for the low percentage of agreement on the ratings may have been due to the timing of the observations and the use of an administrator and peer raters. Administrators tended to be more lenient in rating the teachers than the peer raters. When the raters observed the same classroom lesson, the reliability increased to .94.

The other independent variable was teaching experience that was measured as the number of years teaching. The teachers were categorized into one of the following groups:

- new teacher; 0-3 years
- level I experienced teacher; 4-12 years of experience
- level II experienced teacher; exceeds 12 years of experience

While much of the research affirms that experienced teachers typically are more effective than beginning teachers, no definite categories of teaching experience have been found that provide specific classifications of novice and veteran teachers (Kane, Rockoff, Staiger, 2006).

The dependent variables were the MEAP mathematics and MEAP English language arts assessment. The scaled scores were categorized into performance level

ranges. The third grade performance ranges were numerically coded with numbers ranging from 189 to 439. The state of Michigan assigns a performance level of advanced, proficient, partially proficient, or not proficient to the numerical total for each of the tests. Table 9 shows how the scaled scores for the student MEAP data were calculated to obtain a classroom average for each teacher.

Table 9

Sample of Calculation of Scaled Score to Class Average

Teacher	Yrs. Exp.	Student #	Reading Score	Perf. Rating
1827	24	3252	341	Highly effective
1827	24	4142	350	Highly effective
1827	24	3811	361	Highly effective
1827	24	3429	350	Highly effective
1827	24	4162	327	Highly effective
1827	24	4898	345	Highly effective
1827	24	5371	361	Highly effective
1827	24	6893	350	Highly effective
	Class average reading		344.04	

In 2005 the Office of Assessment and Accountability OEAA reported "The MEAP tests have been recognized nationally as sound, reliable, and valid measurements of academic achievement" (pg. A-2). From empirical item response theory (IRT) used to determine reliability among subgroups, the BAA concluded that the MEAP tests are reliable for different subgroups, such as gender, socioeconomic status, ethnicity, and students with limited English proficiency. Subject and grade level reliability studies supported these conclusions (MDE, BAA, & Measurement Incorporated, 2011-2012). Table 10 presents the reliability statistics for the MEAP assessment. The classification

accuracy was also examined on the MEAP assessments based on the cut scores recommended by the educator panels. Classification accuracy is computed through an IRT model using ability scores, observed scores, and the mean of the standard deviation of the standard of error.

Table 10

Summary Reliability Statistics of Coefficient Alphas Across Subjects and Grade Levels

Subject	Grade	Low	Middle	High
Mathematics	3	.88	.90	.89
Mathematics	6	.88	.91	.89
English Language Arts	3	.84	.85	.84
English Language Arts	6	.86	.86	.86

Note. Adapted from *MEAP Technical Report* by *(MDE)*, 2011-2012, Retrieved from http://www.michigan.gov/documents/mde/MEAP_2010-2011_Technical_Report_394693_7.pdf

The FFT has been found to be both valid and reliable in measuring teacher performance. The MEAP assessment is both valid and reliable as a criterion-based test used to measure student achievement based on state standards. Years of experience is a demographic variable that cannot be tested for validity or reliability, although the information for this variable has been obtained from official school records.

Data Collection Procedures

In October of 2014, the researcher was granted permission from the superintendents in both school districts to access the data needed for the study. The data collection approval documents from the districts' superintendents are included in Appendix A. Approval to conduct the study through the Baker University Institutional Review Board (IRB) was granted on May 22, 2015 (see Appendix B). After IRB

approval, the data processing administrator and the administrator responsible for providing the evaluation scores at the two school districts coded the names and were responsible for integrating the MEAP and teacher data. The third and sixth-grade fall English language arts and mathematics MEAP average scaled scores for each second and fifth-grade teacher in the two school districts for the 2012-2013 and the 2013-2014 academic years were obtained from the school districts' records by the district administrator in charge of data processing. To ensure confidentiality among the teachers, all identifying information was redacted from the data. No additional data were collected from students or teachers. The student results for the MEAP were matched with the teacher from the previous year who taught the students the content for the MEAP. The rationale behind matching grade level teacher evaluations with the following year's results was that the MEAP test is given in the fall, six weeks after the start of the school year, and, therefore, represents what students learned in the previous year. Second-grade teacher evaluation ratings were matched with third-grade MEAP test results, and fifth-grade teacher evaluation ratings were matched with sixth-grade student results.

Data Analysis Methods

The data were provided in an Excel® spreadsheet that included the rating highly effective or effective for each teacher, grade level taught, and years of experience. The researcher computed the averages for the reading and mathematics MEAP results for each teacher. The Excel® data from the data administrator was exported into IBM® SPSS® Statistical Faculty Pack 23 for Windows for analysis. The data were summarized using frequency distributions to provide information on the number of teachers with

evaluation ratings of highly effective and effective and their experiences as new, level I, and level II.

The four research questions were addressed using 2 x 3 factorial analysis of variance. The main effect measures the difference in mean scores between levels of the variable (e.g., highly effective and effective teachers). The interaction is a measure of the difference between variables. Differences were determined using two-way analysis of variance (ANOVA). The two factors were teacher performance (highly effective and effective) and experience (new, level I and level II). Separate 2 x 3 ANOVAs were used to determine if third-grade mathematics results and English language arts results differed for each of the two main effects, teacher performance as measured by teacher evaluations of highly effective and effective and years of experience divided into three categories, new teachers, level I, and level II. This same analysis will be repeated for sixth-grade mathematics and English language arts results. The interaction effect of teacher performance and years of experience also will be tested. If the difference is not statistically significant between effective and highly effective teachers on the third-grade mathematics performance, no further intervention would be needed. If the difference were statistically significant, then the group that was found to have lower scores would be invited to participate in professional development programs to help them develop strategies to improve how they teach mathematics.

If a statistically significant result was obtained for teacher performance, the mean scores for teacher performance were examined to determine the direction of the difference between effective and highly effective teachers. If the results for years of experience were statistically significant, Scheffé post hoc tests were used to compare all

possible pairwise comparisons to determine which of the three levels were contributing to the statistically significant result. If the interaction effect is statistically significant, simple effects were used to determine which groups were contributing to the statistically significant difference in sixth-grade student average fall English language arts performance. Simple effects would examine the effect of one level of an independent variable against the second independent variable. In the present study, teacher performance would be divided into two levels: highly effective and effective. The three levels of teacher experience (new, level I and level II) would be compared to each level of teacher performance separately to determine which level of teacher experience is contributing to the statistically significant interaction effect. All decisions on the statistical significance of the findings will be made using an alpha level of .05.

RQ1. Does second-grade teacher experience, teacher effectiveness, and the interaction of experience and effectiveness influence student mathematics achievement in the third-grade when measured with the fall MEAP?

H1. Second-grade teachers' performance and experience had a significant impact on third-grade student mathematics performance as measured by the fall MEAP assessment. A factorial research design will be conducted using the two-way ANOVA analysis to challenge the hypothesis. An alpha level of significance will be set at .05.

RQ2. Does second-grade teacher experience, teacher effectiveness, and the interaction of experience and effectiveness influence student English language arts achievement in the third-grade when measured with the fall MEAP?

H2. Second-grade teachers' performance and experience had a significant impact on third-grade student English language arts performance as measured by the fall MEAP

assessment. A factorial research design will be conducted using the two-way ANOVA analysis to challenge the hypothesis. An alpha level of significance will be set at .05.

RQ3. Does fifth-grade teacher experience, teacher effectiveness, and the interaction of experience and effectiveness influence student mathematics achievement in the sixth-grade when measured with the fall MEAP?

H3. Fifth-grade teachers' performance and experience had a significant impact on sixth-grade student mathematics performance as measured by the fall MEAP assessment. A factorial research design will be conducted using the two-way ANOVA analysis to challenge the hypothesis. An alpha level of significance will be set at .05.

RQ4. Does fifth-grade teacher experience, teacher effectiveness, and the interaction of experience and effectiveness influence student English language arts achievement in sixth-grade when measured with the fall MEAP?

***H4*.** Fifth-grade teachers' performance and experience had a significant impact on sixth-grade student English language arts performance as measured by the fall MEAP assessment. A factorial research design will be conducted using the two-way ANOVA analysis to challenge the hypothesis. An alpha level of significance will be set at .05.

Limitations

According to Lunenburg and Irby (2008, p. 133), "limitations are factors that may have an effect on the interpretation of the findings or on the generalizability of the results." Though accurate data collection is essential for any study and leads to precise conclusions limitations exist and are not under the researcher's control. Lunenburg and Irby recommend that readers should be given insight regarding limitations, which may

influence the results of the study (Lunenburg and Irby, 2008). Limitations for this study included:

1. The nature of evaluation is subjective even when evaluators are provided with training focused on a specific set of criteria. Different evaluators conducted the observations and evaluations, resulting in subjectivity. This subjectivity may have limited the results of the study.

2. Since the MEAP assessment was given six weeks after the start of the school year, it measures learning from the previous school year. However, summer learning loss is not considered. Unless students participated in summer school or other summer learning programs the absence of schooling in the summer can lead to a loss of learning.

3. Summer learning regression was not factored into the study.

4. Inconsistencies exist from classroom to classroom and school to school regarding classroom expectations, staff culture, instructional strategies, preparation for testing, student behavior and testing environment. The result of the inconsistencies may affect the outcome.

Summary

The research methodology and procedures were described in chapter three and used to evaluate the impact that teachers' evaluation scores had on student achievement. A non-experimental, ex-post facto research design was utilized in the study. The research design, population and sample, sampling procedures, measurement, validity and reliability, description of the Michigan teacher evaluation samples, the data collection procedures, the data analysis and hypothesis testing, and limitations of the study were presented. The results of the current study are presented in chapter four.

Chapter Four

Findings

The results of the statistical analyses that were used to address the four research questions posed for this study are presented in this chapter. The data for reading and mathematics scores from the Michigan Education Assessment Program (MEAP) were obtained from closed school records for the two academic years (2012-2013 and 2013-2014) and were used as the dependent variables in this study. The independent variables were teacher evaluation (highly effective or effective) and years of teaching experience, new, level I, and level II.

The purpose of this study was to investigate the impact that teacher performance and experience had on student performance in mathematics and English language arts, as measured by the MEAP assessment and teachers' evaluation ratings. The data used in this study were obtained from a sample of students in grades three and six, and teachers (n= 96) who taught second and fifth grades. An additional purpose was to determine if there was a difference between teachers' years of experience and student English language arts and mathematic MEAP results. This chapter presents the descriptive statistics used for analysis and results of the data analysis for the research questions.

Descriptive Statistics

The teacher evaluation ratings and the number of years of experience were summarized by grade level from the two school districts. The data were summarized using frequency distributions. Table 11 presents the summarized results of these analyses for second grade teachers.

The largest group of 2nd grade teachers were level I for both 2012-13

(*n* = 48.9%) and 2013-14 (*n* = 44.7%). Twenty-six (100.0%) teachers in 2012-13 were rated as highly effective compared to 36 (100.0%) teachers rated highly effective in 2013-14. In 2012-13, fewer new teachers were rated highly effective (*n* = .5%) than effective (*n* = 33.3%). In contrast, more new teachers were rated highly effective (*n* = 13.9%) than effective (*n* = 18.2%) in 2013-14.

Table 11

Demographic Characteristics: 2nd Grade Teacher Effectiveness and Years of Experience

	Teacher Effectiveness					
	Highly Effective		Effective		Total	
Years of Experience	*n*	%	*n*	%	*n*	%
2012-13						
New	3	11.5	7	33.3	10	21.3
Level I	15	57.7	8	38.1	23	48.9
Level II	8	30.8	6	28.6	14	29.8
Total	26	100.0	21	100.0	47	100.0
2013-14						
New	5	13.9	2	18.2	7	14.9
Level I	15	41.7	6	54.5	21	44.7
Level II	16	44.4	3	27.3	19	40.4
Total	36	100.0	11	100.0	47	100.0

The years of experience for the 5th grade teachers were cross-tabulated by their evaluation ratings of either highly effective or effective. As shown in Table 12, fewer 5th grade teachers in 2012-13 (*n* = 100.0%) were rated highly effective than in 2013-14 (*n* = 100.0%). Level II teachers were more likely to be rated highly effective in both 2012-13 (*n* = 41.9%) than in 2013-14 (*n* = 38.8%). New teachers in 2012-13 (*n* = 26.1%) were rated highly effective, with 9 (24.3%) new teachers rated highly effective in 2013-14.

Table 12

Demographic Characteristics: 5th Grade Teacher Effectiveness and Years of Experience

	Teacher Effectiveness					
	Highly Effective		Effective		Total	
Years of Experience	*n*	%	*n*	%	*n*	%
2012-13						
New	6	26.1	3	15.0	9	20.9
Level I	11	47.8	5	25.0	16	37.2
Level II	6	26.1	12	60.0	18	41.9
Total	23	100.0	20	100.0	43	100.0
2013-14						
New	9	24.3	5	41.7	14	28.6
Level I	13	35.2	3	25.0	16	32.6
Level II	15	40.5	4	33.3	19	38.8
Total	37	100.0	12	100.0	49	100.0

Hypothesis Testing

Four research questions and associated hypotheses were developed for the study. These questions were addressed and the hypotheses tested using multivariate analysis of variance. All decisions on the statistical significance of the findings were made using a criterion alpha level of .05.

RQ1. Does second-grade teacher experience, teacher effectiveness, and the interaction of experience and effectiveness influence student mathematics achievement in the third-grade when measured with the fall MEAP?

H1. Second-grade teacher performance and experience had a significant impact on third-grade student mathematics performance as measured by the fall MEAP assessment. A factorial research design was conducted using the two-way ANOVA analysis to challenge the hypothesis. An alpha level of significance was set at .05.

A 2 x 3 factorial ANOVA was used to test the hypotheses that second grade

teacher performance and experience had a significant impact on third-grade student mathematics performance on the Fall MEAP assessment. The dependent variable in this analysis was third grade mathematics MEAP results. The independent variables were the evaluation ratings as either highly effective or effective and the years of experience, new, level I, and level II. Table 13 presents the descriptive statistics for the mathematics scores.

Table 13

Descriptive Statistics: Mathematics Scores for Third Grade MEAP by Teacher Effectiveness Ratings and Years of Experience

Independent Variable	*n*	*M*	*SD*
Teacher Effectiveness Ratings			
Highly Effective	62	346.34	12.50
Effective	32	346.38	27.43
Years of Experience			
New – 1 to 3 years	17	350.70	13.90
Level I – 4 to 12 years	44	346.41	18.44
Level II – Over 12 years	33	344.03	21.39
Teacher Effectiveness Ratings X Years of Experience			
Highly effective x New – 1 to 3 years	8	345.13	15.47
Highly effective x Level I – 4 to 12 years	30	347.25	12.18
Highly effective x Level II – Over 12 years	24	345.60	12.33
Effective x New – 1 to 3 years	9	355.64	10.88
Effective x Level I – 4 to 12 years	14	344.60	28.08
Effective x Level II – Over 12 years	9	339.87	36.95

The results of the 2 x 3 factorial analysis of variance for third grade mathematics MEAP results by 2nd grade teacher effectiveness and years of experience are presented in Table 14. The results of the analysis comparing the third grade mathematics MEAP results by teacher effectiveness and years of experience did not provide any evidence of statistically significant differences (F (1, 88) = .03, p = .873) or years of experience (F (2, 88) = .84, p = .435). The interaction effect between teacher effectiveness and years of experience was not statistically significant, F (2, 88) = 1.02, p = .365. Due to the lack of

significance for H1, no mean comparisons were made. The hypothesis was not supported.

Table 14

Between Subjects Effects: Mathematics Scores for Third Grade MEAP by Teacher Effectiveness Ratings and Years of Experience

Source	Sum of Squares	Mean Squares	*df*	*F*	*p*
Effectiveness Rating	9.23	9.23	1	.03	.873
Years of Experience	603.61	301.84	2	.84	.435
Effectiveness Rating X Years of Experience	732.74	366.37	2	1.02	.365
Error	31602.12	359.12	88		

RQ2. Does second-grade teacher experience, teacher effectiveness, and the interaction of experience and effectiveness influence student English language arts achievement in the third-grade when measured with the fall MEAP?

H2. Second-grade teacher performance and experience had a significant impact on third-grade student English language arts performance as measured by the fall MEAP assessment. A factorial research design was conducted using the two-way ANOVA analysis to challenge the hypothesis. An alpha level of significance was set at .05.

A 2 x 3 factorial ANOVA was used to test the hypotheses that second-grade teacher performance and experience had a significant impact on third-grade student English language arts performance on the fall MEAP assessment. The dependent variable in these analyses was third-grade English language arts MEAP results. The independent variables were the evaluation ratings as either highly effective or effective and the years of experience, new, level I, and level II. Table 15 presents the descriptive statistics for the English language arts scores.

Table 15

Descriptive Statistics: English Language Arts Scores for Third Grade MEAP by Teacher Effectiveness Ratings and Years of Experience

Independent Variable	*n*	*M*	*SD*
English Language Arts			
Teacher Effectiveness Ratings			
Highly Effective	62	342.93	14.08
Effective	32	345.95	11.97
Years of Experience			
New – 1 to 3 years	17	343.15	10.19
Level I – 4 to 12 years	44	342.89	15.01
Level II – Over 12 years	33	345.80	12.76
Teacher Effectiveness Ratings X Years of Experience			
Highly effective x New – 1 to 3 years	8	339.52	10.01
Highly effective x Level I – 4 to 12 years	30	342.70	15.68
Highly effective x Level II – Over 12 years	24	344.36	13.39
Effective x New – 1 to 3 years	9	346.39	9.75
Effective x Level I – 4 to 12 years	14	343.29	14.02
Effective x Level II – Over 12 years	9	349.65	10.61

The main effects of effectiveness ratings and years of experience used to compare third-grade English language arts scores were not statistically significant. The interaction effect between effectiveness ratings and years of experience was not statistically significant. When the third-grade English language arts scores were compared between teachers rated as highly effective and effective, the result was not statistically significant, F (1, 88) = 1.79, p = .184. Due to the lack of significance for H2, no mean comparisons were made. As a result of these findings, there was no significant difference in third grade mathematics scores by teacher effectiveness and years of experience therefore the hypothesis was not supported.

Table 16

Between Subjects Effects: *English Language Arts Scores for Third Grade MEAP by Teacher Effectiveness Ratings and Years of Experience*

Source	Sum of Squares	Mean Squares	*df*	*F*	*p*
Effectiveness Rating	329.00	329.00	1	1.79	.184
Years of Experience	287.06	143.53	2	.78	.461
Effectiveness Rating X Years of Experience	149.46	74.73	2	.41	.667
Error	16169.01	183.74	88		

RQ3. Does fifth-grade teacher experience, teacher effectiveness, and the interaction of experience and effectiveness influence student mathematics achievement in the sixth-grade when measured with the fall MEAP?

H3. Fifth-grade teacher performance and experience had a significant impact on sixth-grade student mathematics performance as measured by the fall MEAP assessment. A factorial research design was conducted using the two-way ANOVA analysis to challenge the hypothesis. An alpha level of significance was set at .05.

A 2 x 3 factorial ANOVA was used to test the hypotheses that fifth-grade teachers' performance and experience had a significant impact on sixth-grade student mathematics performance on the Fall MEAP assessment. The dependent variable in this analysis was sixth-grade mathematics MEAP results. The independent variables were the evaluation ratings as either highly effective or effective and the years of experience, new, level I, and level II. Table 17 presents the descriptive statistics for the mathematics scores.

Table 17

Descriptive Statistics: Mathematics Scores for Sixth-Grade MEAP by Teacher Effectiveness Ratings and Years of Experience

Independent Variable	*N*	*M*	*SD*
Teacher Effectiveness Ratings			
Highly Effective	60	589.76	63.01
Effective	32	618.38	61.12
Years of Experience			
New – 1 to 3 years	23	591.64	57.50
Level I – 4 to 12 years	32	609.58	60.85
Level II – Over 12 years	37	596.22	69.64
Teacher Effectiveness Ratings X Years of Experience			
Highly effective x New – 1 to 3 years	15	594.22	58.40
Highly effective x Level I – 4 to 12 years	24	606.40	58.19
Highly effective x Level II – Over 12 years	21	567.57	67.60
Effective x New – 1 to 3 years	8	586.80	59.39
Effective x Level I – 4 to 12 years	8	619.12	71.67
Effective x Level II – Over 12 years	16	633.80	53.77

The results of the 2 x 3 factorial analysis of variance presented in Table 18 provided no evidence of statistically significant differences for either main effect, effectiveness ratings, or years of experience. The comparison of mathematics scores for sixth-grade MEAP did not result in differences between teachers rated highly effective or effective, F (1, 86) = 2.93, p = .091. Due to the lack of significance for H3, no mean comparisons were made. The hypothesis was not supported.

Table 18

Between Subjects Effects: Mathematics Scores for Sixth-Grade MEAP by Teacher Effectiveness Ratings and Years of Experience

Source	Sum of Squares	Mean Squares	*df*	*F*	*p*
Effectiveness Rating	10924.11	10924.11	1	2.93	.091
Years of Experience	5586.23	2793.11	2	.75	.476
Effectiveness Rating X Years of Experience	20974.52	10487.26	2	2.81	.066
Error	321028.10	3732.89	86		

RQ4. Does fifth-grade teacher experience, teacher effectiveness, and the interaction of experience and effectiveness influence student English language arts achievement in sixth-grade when measured with the fall MEAP?

H4. Fifth-grade teacher performance and experience had a significant impact on sixth-grade student English language arts performance as measured by the fall MEAP assessment. A factorial research design was conducted using the two-way ANOVA analysis to challenge the hypothesis. An alpha level of significance was set at .05.

A 2 x 3 factorial analysis of variance was used to compare the sixth-grade English language arts outcomes by teacher effectiveness ratings and years of experience. The dependent variable in this analysis was sixth grade English language arts scores on the MEAP tests for the 2012-13 and 2013-14 academic years. The independent variables were the teacher effectiveness ratings (highly effective and effective) and years of experience. Results of the descriptive statistics for this analysis are presented in Table 19. A statistically significant difference was found for sixth-grade English language arts MEAP scores for the main effect of teacher effectiveness ratings, F (1, 86) = 4.16, p=.044. The mean scores for teachers rated effective (M = 612.45, SD = 52.56) were

higher than for teachers who were rated highly effective (M = 585.23, SD = 48.65). The sixth-grade English language arts scores did not differ among teachers relative to their years of experience, F (2, 86) = 2.93, p = .091. The teachers with 4 to 12 years of experience (M = 597.73, SD = 52.59) and those with more than 12 years of experience (M = 597.52, SD = 52.93) were higher than scores for teachers with 0 to 3 years of experience (M = 585.00,

SD = 49.47). The interaction effect of effectiveness ratings by years of experience was not statistically significant, F (2, 86) = 1.66, p = .197). The highest mean scores were on the sixth-grade English language arts MEAP test and were obtained by teachers rated effective with over 12 years of experience (M = 625.68, SD = 47.54), while the lowest scores were found for teachers rated highly effective with more than 12 years of experience (M = 576.07, SD = 47.22). Based on the mixed findings for this analysis, H4 was supported.

Table 19

Descriptive Statistics: English Language Arts Scores for Sixth-Grade MEAP by Teacher Effectiveness Ratings and Years of Experience

Independent Variable	*n*	*M*	*SD*
Teacher Effectiveness Ratings			
Highly Effective	60	585.23	48.65
Effective	32	612.45	52.56
Years of Experience			
New – 1 to 3 years	23	585.94	48.54
Level I – 4 to 12 years	32	597.73	52.59
Level II – Over 12 years	37	597.52	52.93
Teacher Effectiveness Ratings X Years of Experience			
Highly effective x New – 1 to 3 years	15	585.00	49.47
Highly effective x Level I – 4 to 12 years	24	593.39	49.97
Highly effective x Level II – Over 12 years	21	576.07	47.22
Effective x New – 1 to 3 years	8	587.72	50.06
Effective x Level I – 4 to 12 years	8	610.75	61.53
Effective x Level II – Over 12 years	16	625.68	47.54

As shown on Table 20, a statistically significant difference was found on sixth-grade English language arts MEAP scores for the main effect of teacher effectiveness ratings, F (1, 86) = 4.16, p = .044. The mean scores for teachers rated effective (M = 612.45, SD = 52.56) were significantly higher than for teachers who were rated highly effective (M = 585.23, SD =48.65). The sixth-grade English language arts scores did not differ among teachers relative to their years of experience, F (2, 86) = 2.93, p = .091. The teacher evaluation rating with 4 to 12 years of experience (M = 597.73, SD = 52.59) and those with more than 12 years of experience (M = 597.52, SD = 52.93) were higher than scores for teachers with 0 to 3 years of experience (M = 585.00, SD = 49.47).

The interaction effect of effectiveness ratings by years of experience was not statistically significant, F (2, 86) = 1.66, *p* = .197). The highest mean scores were on the sixth-grade English language arts MEAP test and were obtained by teachers rated effective with over 12 years of experience (M = 625.68, SD = 47.54), while the lowest scores were found for teachers rated highly effective with more than 12 years of experience (M = 576.07, SD = 47.22). Based on the mixed findings for this analysis, H4 was supported.

Table 20

Between Subjects Effects: *English Language Arts Scores for Sixth-Grade MEAP by Teacher Effectiveness Ratings and Years of Experience*

Source	Sum of Squares	Mean Squares	*df*	*F*	*p*
Effectiveness Rating	10367.24	10367.24	1	4.16	.044
Years of Experience	3504.07	1752.04	2	2.93	.091
Effectiveness Rating X Years of Experience	8251.31	4125.65	2	1.66	.197
Error	214215.71	2490.88	86		

Summary

Chapter four has presented the results of the data analysis used to describe the sample and test the hypotheses. Descriptive statistics were used to provide information on the teacher effectiveness ratings and the mean scores for the fourth and sixth grade English language arts and mathematics MEAP tests. Factorial analyses of variance were used to test the four hypotheses and address the research questions posed for the study. A discussion of the findings, conclusions, and implications are presented in chapter five.

Chapter Five

Interpretation and Recommendations

Accountability through standardized testing has been a trend before 2001 when No Child Left Behind (NCLB) emerged. Across the nation, teacher evaluation models are being revised to include student achievement growth criteria. This study examined teacher evaluation ratings and years of experience to determine if these factors impacted student achievement. In chapter five the researcher summarizes the study providing an overview of the problem, the purpose, and the research questions. The methodology is reviewed, as well as the major findings from the research. The study concludes with recommendations for educators and suggestions for further research studies related to the topic.

Overview of the Problem

The focus of this study was to determine the impact of teacher evaluation and teacher experience on student achievement data as measured by the Michigan Educational Assessment Program (MEAP). The results of the current study did not indicate a significant relationship existed between teacher evaluation ratings and experience and student achievement. As indicated in chapter one, no published literature has been found in Michigan that supports a link between student performance and teacher effectiveness ratings. However, mandated by the state, school districts are required to have, as one component of the evaluation tool, a student growth measure. The state of Michigan also requires districts to establish performance-based pay incentives for teachers who earn a highly effective rating. However, in establishing criteria for performance-based pay incentives, inconsistencies exist between different schools and

different teachers within a school in determining student growth. Since student growth is weighted heavily in the overall evaluation, the districts in the study are investigating consistent ways to measure student growth. Using standardized testing data is a consideration. If valid, using the same test data from all schools within a district and within the state could be instrumental in minimizing subjectivity. The intent of this study was to examine the link between teacher effectiveness and years of experience related to student achievement.

Purpose Statement and Research Questions.

The purpose of this study was to determine if teacher performance, as measured by teacher final evaluation ratings, had an impact on elementary students' academic achievement. The second purpose of the study was to determine if teacher years of experience had an impact on student academic performance. Additionally, the study examined the interaction between teacher experience and teacher effectiveness, and its impact on student achievement. Specifically, the intent of the current study was to identify if second and fifth-grade teacher evaluation ratings and years of experience made an impact on student achievement as measured by the third and sixth-grade fall MEAP assessment for mathematics and English language arts.

Review of the Methodology. A non-experimental, ex-post facto research design was used to study second grade and fifth-grade teachers from two suburban Detroit school districts. Only highly effective and effective second and fifth grade teacher evaluation data were used. Data were analyzed for the 2012-2013 and 2013-2014 school years. Student data from third and sixth-grade fall mathematics and English language arts MEAP scores were provided by both districts' data processing departments. The

averages for the reading and mathematics MEAP results were compiled for each teacher. A 2 x 3 factorial ANOVA was used to address the four research questions. The two factors were teacher performance (highly effective and effective) and experience (new, level I, and level II). The interaction of performance and years of experience was also tested.

Major findings. Data were obtained on teacher effectiveness ratings for 47 second-grade teachers and 49 fifth grade teachers. Teachers were rated as either highly effective or effective. The years of experience were categorized into three groups, new (1 to 3 years), level I (4 to 12 years), and level II (exceeding12 years). Four research questions and hypotheses were developed for the study. Hypothesis one used a 2 x 3 factorial ANOVA to compare third-grade mathematics scores on the MEAP by teacher effectiveness ratings and years of experience. The hypothesis was not supported as the findings were not significant. The second hypothesis used a 2 x 3 factorial ANOVA to determine if third-grade English language arts scores on the MEAP by teacher effectiveness ratings and years of experience. The results of this analysis were not statistically significant, indicating the hypothesis was not supported. The sixth-grade mathematics scores on the MEAP were used as the dependent variable in a 2 x 3 ANOVA with teacher effectiveness ratings and years of experience used as the independent variables to test the third hypothesis. The results of this analysis were not statistically significant, indicating the hypothesis was not supported. The fourth hypothesis was used to compare the sixth-grade English language arts scores on the MEAP by teacher effectiveness and years of experience. The English language arts scores differed for teacher effectiveness ratings, with teachers who were rated effective

having higher mean scores than teachers who were rated as highly effective. No statistically significant differences were found for years of experience or the interaction between teacher effectiveness ratings and years of experience. The results of this analysis were not statistically significant, indicating the hypothesis was not supported.

The current study revealed that years of experience, teacher effectiveness ratings, and the interaction between the two did not have an influence on student performance on the fall MEAP assessment. There was no impact from teacher experience on student performance on the fall MEAP tests in math and English language arts. Results from the findings of the current research revealed that teacher experience did not have an impact on student achievement in third grade or sixth-grade on the fall MEAP for mathematics. There was no impact from teacher effectiveness shown in the findings for English language arts for third grade. The data did not support an interaction effect of teacher effectiveness and student achievement for either mathematics or English language arts fall MEAP tests.

These findings were true for both third-grade and sixth-grade over the two academic years included in the study. This study did not find that second grade teacher experience had an impact on students who took the third grade math and English language arts MEAP test in the fall. To summarize, this study revealed that student achievement did not differ based on years of experience, teacher performance ratings, or the interaction between the two.

Findings Related to the Literature

The current study focused on teacher end of the year performance ratings, teacher years of experience, and the interaction of the two and whether or not these factors made

an impact on student academic achievement. The goal of this study was to add to the body of research correlating teacher performance and student achievement. Each of the four hypotheses predicted that teacher experience and performance, and the interaction between the two had a significant impact on student performance.

Though the majority of principals, parents, and students believed that teacher quality equated to student achievement, minimal evidence has been found that supports this notion (Hanushek, 1986). The results of the current study provided evidence that student achievement is not influenced by teacher quality. The researcher found that teacher effectiveness did not impact student MEAP scores significantly for either mathematics or English language arts in third-grade. Linking educator quality to student performance is a way for schools to be held accountable and focus on results (Braun, 2005). This study did not support Braun's statement. The results of the current study revealed teacher effectiveness had no impact on student scores for the sixth-grade MEAP mathematics test. The only statistically significant difference identified by the results was a statistically significant difference between teachers rated highly effective or effective on sixth-grade English language arts MEAP scores. The mean student performance scores for the sixth-grade MEAP English language arts test were higher for effective teachers than highly effective teachers. The findings from the current study did not indicate that teacher quality or performance influenced student achievement. Furthermore, credentials that teachers earn do not always equate to quality classroom instruction (Toch & Rothman, 2008).

Rockoff (2004) argued that teacher quality consisted of observable characteristics that could not be measured. The Danielson Framework for Teaching (FFT) model is

based on classroom teaching observations that are focused on descriptors of quality teaching (Danielson, 2007). Weigberg et al. (2009) determined that quality teachers had a positive impact on student achievement and success, but the concern was that too many teachers were rated at the top level. The findings of the current study were not consistent with those of Weigberg et al. Classroom observation is one area used in the overall annual teacher evaluation. Student performance also is factored into teacher overall performance. A teacher can be found to be highly effective in instructional delivery, but could still have less than proficient student achievement results. Several studies determined that teacher effectiveness is the most important factor that influences student success (Bill and Melinda Gates Foundation, 2013; Tucker and Stronge, 2005). Results from the current study did not find this statement to be true.

Hanusek and Rivkin, 2003 did not identify a direct link between teacher experience to teacher quality. Other studies indicated that educators become more competent with experience (Danielson, 2010; Greenwald, Hedges, & Laine, 1996; Hanushek, Kain, & Rivkin, 2005; Kane, Rockoff, & Staiger, 2006). According to Greenwald, Hedges, & Laine, studies on the effect of years of experience on student outcomes indicated that experienced teachers produced higher student test scores (Greenwald, Hedges & Laine, 1996). Similar findings were not found in the present study. Students of experienced teachers had higher achievement scores than students of teachers with less than three years of experience. First-year teachers with higher student achievement gain tended to produce even greater gains in year two (Rockoff & Speroni, 2011). None of the results from the current study indicated that teachers' years of experience had an impact on MEAP scores for third or sixth-grade students.

The focus of the research questions was to determine if an interaction between teacher experience and teacher effectiveness influenced student academic achievement. The researcher did not find any literature or studies, which encompassed the interaction between years of experience and teacher performance and how these impact student achievement. The study results revealed no significant correlation between years of experience, performance ratings, and the interaction between the two on student achievement.

Conclusions

This study provided results regarding teacher effectiveness and student academic performance growth. School leaders should recognize that teacher effectiveness ratings do not impact student performance. School administrators also should be aware that years of experience might not result in higher student achievement. Student achievement is not dependent on the interaction between teacher effectiveness and on years of experience. Regardless of teacher experience and effectiveness, the students can be expected to perform at an equal level. Results of this study provided little evidence to support a link between student performance and evaluation of teacher performance or experience. Implications for action and future research are included in the next section of this study.

Implications for Action. The research results should be used to influence teacher evaluation mandates that are based in large part on student achievement. The current study did not show that student achievement is impacted by teacher performance. Legislators should consider this body of work and reduce the weight of the student growth component on teacher evaluations. Teachers could use this study to advocate for

change in the current teacher evaluation policy. As data from this research revealed that teacher performance did not directly support student performance, the study results could open a dialogue to determine meaningful ways to measure student academic growth and provide much-needed consistency between and within school districts.

Student growth measures have the highest weight of the overall evaluation in determining teachers who will be considered either highly effective or effective. The measures used to determine student growth need to be consistent throughout the state, and especially throughout districts and schools. The most current legislation for the state of Michigan indicates that starting in 2018-2019 academic year, 40% of the annual year-end teacher evaluation must be based on student achievement growth. Core content areas will be required to have 50% of the student growth come from state assessments. Policy makers need guidance from expert teachers to ensure that meaningful learning is a priority. When schools have significantly different expectations for student achievement growth measures, overall teacher evaluation ratings could be impacted. A highly effective teacher in one building or district may be an effective teacher in another building or district. Accountability for student growth is advisable, however if the approach is not solid, it can become a practice of compliance void of substance.

Highly effective teachers should be in every classroom. This particular study did not indicate that years of experience significantly influenced achievement positively or negatively. So whether a teacher is new, level I, or level II, clear expectations of what is required to be highly effective should be articulated to all teachers. If teachers know the criteria required to be considered highly effective, and the instructional leaders within districts and buildings provide time for teachers to enhance and develop their skills

around these criteria, the number of highly effective teachers in all classrooms is likely to increase.

Consideration should be given to providing opportunities for multiple evaluators or peer evaluation. Perhaps the highly effective teachers could be part of the evaluation process. This could be a powerful growth measure and support, if a trusting culture is in place, for both student and teacher development. Multiple evaluators could also ensure inter-rater reliability within the evaluation process.

Recommendations for future research. A longitudinal study could be conducted to examine standardized test outcomes for cohorts of students for a period of three years or more. The purpose of this research is to determine the impact of teacher ratings on student outcomes over time. Students would be sorted according to who had highly effective and effective teachers over time. The study could determine if students who had highly effective teachers over three years had higher achievement scores. Multiple assessments over time, both formative and summative, versus one single state assessment, should be considered as the dependent variables in this study. If a statistically significant difference is found in academic achievement between students whose teachers are consistently rated as highly effective and those who are being taught by teachers with inconsistent ratings, policy makers could work to improve professional development for teachers who are not rated as highly effective.

In order to influence Michigan policy makers to rethink their decisions to focus exclusively on standardized tests to measure teacher accountability for student achievement growth further studies need to be conducted. Instead of using one single assessment, policy makers should consider the use of portfolio assessments to monitor

student achievement progress throughout the year. The principal or evaluator could examine the artifacts and determine if students are learning from the evidence provided in the portfolios. This type of authentic assessment of student learning would be a better way to evaluate the teachers' effectiveness in the classroom.

Additional research should focus on the effect of mentoring programs for new teachers. Many school districts promote a mentoring program for new teachers, but do little to monitor the outcomes. A mentoring program can be useful, but only if the mentor-mentee are actively involved.

As districts across the country attempt to keep up with new mandates from the federal and state legislature regarding teacher evaluation, school officials need to be informed and aware of meaningful ways to ensure consistent and fair evaluations. Conducting a study to obtain perceptions of school administrators on how teachers are evaluated in their school districts could provide useful information regarding the consistency of teacher ratings. The results of this study could be used by state boards of education to make teacher evaluation mandates more relevant to student outcomes.

Replicating this study with more school districts across the state of Michigan could lead to different results due to the lack of consistency in teacher evaluations. This study could include teachers at all levels of the teacher evaluation; highly effective, effective, minimally effective, and minimally effective. Both of the schools in this study used the Danielson Framework for Teaching (FFT) evaluation model. It would be interesting to see if the results would be similar for other evaluation models such as the Marzano model or the 5 Dimensions of Teaching and Learning Model. Further

information could be drawn from disaggregating student demographics, such as socioeconomic status, English as a Second Language, and gender.

Concluding remarks. Recognizing that teacher accountability is likely to increase, school administrators should be aware of the lack of a link between teacher effectiveness and teacher experience on student achievement. The purpose of the current study was to determine if teacher effectiveness and experience impacted student achievement. There were no statistically significant findings to indicate that teacher effectiveness or experience had an impact on student mathematics or English language arts MEAP performance in third or sixth-grade. The current study results did not align with much of what had been reported in the literature. Suggestions for further research could provide educators with more information on meaningful measurement of teacher performance related to student achievement. Many variables go into learning. Learning is a very complex process and outcomes on standardized tests extend beyond what is taught in the classroom. Hiring experienced teachers over novice ones does not guarantee higher student achievement performance. School officials and state legislators should be aware that teacher evaluation and years of experience is not linked to student achievement.

References

Aiken, L. S., & West, S. G. (1991). *Educational research: Testing and interpreting interactions*. Los Angeles: Sage Publications.

ACT (2014). *ACT technical manual*. Retrieved from http://www.act.org/research/ researchers/index.html

Allemann, J. (2006). *Links between teacher evaluations/supervision an student achievement: a case study of a successful urban elementary school*. (Doctoral dissertation). Retrieved from http://digitallibrary.usc.edu/cdm/compoundobject /collection/p15799coll17/id/109203/rec/29. Available from ProQuest Dissertations (UMI No. 3237169)

Aramath G.A. (2014) *Investigating practices of research-proven multidimensional teacher evaluation system in Michigan schools*. (Doctoral dissertation). Retrieved from scholarworks.wmich.edu/cgi/viewcontent.cgi?article =1235&contex=dissertations.

Barnett, W., Justice, L. M., Sheridan, S. M. (2012). *Handbook of Early Childhood Education.* New York: Guilford Press.

Bherstock-Sherratt, E., Rizzolo, A., Laine, S., & Friedman, W. (2013). *Everyone at the table engaging teachers in evaluation reform*. Hoboken, NJ: John Wiley & Sons Publications

Bill and Melinda Gates Foundation. (2013). Ensuring fair and equitable measures of effective teaching: Culuminating findings from the METs three-year study. Retrieved from http://www.metproject.org/ downloads/MET_Ensuring_ Fair_and_Reliable_Measures_Practitioner_Brief.pdf

Black, P. (1998). *Testing: friend or foe? The theory and practice of assessment and testing*. Washington, D.C.: The Falmer Press.

Braun, H. (2005). *Student progress to evaluate teachers: A primer on value-added models* (Policy Brief). Retrieved from http://files.eric.ed.gov/fulltext /ED529977.pdf

Brophy, J. (1986). Teacher influences on student achievement. *American Psychologist, 41*(10), 1069-1077. doi:10.1037/h0092058

Caliendo, R. J. (1986). Selecting capable teachers. *ERS Spectrum, 4*(1), 22-26.

Center for Educational Leadership (2001). *5 dimensions of teaching and learning*. Retrieved from https://www.k-12leadership.org/research-base

Center for Educational Leadership (2014*). 5 dimensions of teaching and learning*. Retrieved from https://www.k-12leadership.org/content/service/5-dimensions-of-teaching-and-learning.

Center for Educational Performance [Michigan.gov] (2001-2015). Retrieved February 04, 2015, from https://www.mischooldata.org/DistrictSchoolProfiles/ ReportCard/EducationDashboard.aspx

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

Danielson, C. (1996). *Enhancing professional practice: a framework for teaching*. Alexandria VA: Association for Supervision and Curriculum Development

Danielson, C. (2001). New trends in teacher evaluation. *Educational Leadership, 58*(5), 12-15). http://www.ascd.org/publications/educational-leadership.aspx

Danielson, C. (2007). *Enhancing professional practice a framework for teaching* (2nd ed.). Alexandria, VA: ASCD

Danielson, C. (2010). Evaluations that help teachers learn. *Educational Leadership, 68*(4), 35-39. Retrieved from http://www.ascd.org/publications/educational-leadership.aspx

Danielson, C. (2012). Observing classroom practice. *Educational Leadership, 70*(3). Retrieved http://www.ascd.org/publications/educational-leadership.aspx.

Danielson, C. (2013) *Framework for teaching*. Retrieved March 2013, from www.danielsongroup.org.

Danielson, C. & McGreal, T.L. (2000). *Teacher evaluation to enhance professional practice*. Alexandria, VA: Association for Supervision and Curriculum Development. http://www.ascd.org/publications/ educational-leadership.aspx

Darling-Hammond, L. (2010) *Evaluating teacher effectiveness: How teacher performance assessments can measure and improve teaching*. Center for American Progress. Retrieved from http://files.eric.ed.gov/fulltext/ ED535859.pdf

Devine, K. (2009). *Guide to U.S. Department of Education Programs*. Washington D.C.: Diane Publishing Company. Retrieved from https://www2.ed.gov/ programs/gtep/gtep2009.pdf

DiPaola, M. F., Hoy, W. K., & DiPaola, M. F. (2014). *Improving instruction through supervision, evaluation, and professional development*. Charlotte, NC: Information Age Publishing.

"Does Highly Qualified Mean Highly Effective?" Center for Public Education. Nov. 2009. Center for Public Education. Retrieved from http://www.centerforpubliceducation.org/Main-Menu/Staffingstudents/How-good-are-your-teachers-Trying-to-define-teacher-quality/Does-highly-qualified-mean-highly-effective.html

Dufour, R., Dufour, R., Eaker, R., & Many, T. (2006). *Learning by doing*. Bloomington, IN: Solution Tree.

Earl, L. M. (2013). *Assessment as learning: Using classroom assessment to maximize student learning*. Thousand Oaks, CA: Corwin.

Eisner, E.W. (2004). Preparing for today and tomorrow. *Educational Leadership, 61*(4), 6-10. Retrieved from http://www.ascd.org/publications/educational-leadership/dec03/vol61/num04/Preparing-for-Today-and-Tomorrow.aspx

Fink, S., McDermott, J., Austin, S., & Cloninger, K. (2001). *5 dimensions of teaching and learning*. Seattle, WA: Center for Educational Leadership. Retrieved from info.k-12leadership.org/5-dimensions-of-teaching-and-learning.

Fiorina, M. (1989). *Congress: Keystone of the Washington establishment*. Yale University Press.

Fletcher, D. (2009, December 11). Standardized testing. *Time*. Retrieved from http://content.time.com/time/nation/article/0,8599,1947019,00.html

French, R. (2014, February 3). Michigan teachers are under the microscope. *Bridge*. Retrieved from http://bridgemi.com/2014/02/statewide-teacher-evaluation-law-nears-completion/

Garnett, J. (2013) *Teacher effectiveness and experience comparing evaluation ratings and student achievement* (Doctoral dissertation). University of Nebraska. Retrieved from http://files.eric.ed.gov/fulltext/ED521228.pdf. Available from ProQuest Dissertations (UMI No. 1469745865)

Giordano, G. (2005). *How testing came to dominate American schools: The history of educational assessment*. New York: P. Lang.

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

Goldrick, L. (2002). *Improving teacher evaluation to improve teaching quality* (pp. 1-9, Rep. No. ED480159). Washington D.C.: National Governors' Association. Retrieved December, 2002, from eric.ed.gov. (ERIC Document Reproduction Service)

Haertel, E., Ravitch, D., Rothstein, R. (2010, August). *Problems with the use of student test scores to evaluate teachers*. Economic Policy Institute (Briefing Paper No. 278). Retrieved from http://www.epi.org/publication/bp278/

Hazi, M., & Rucinski, D. A. (2009). Teacher evaluation as policy target: Viable reform venue or just another tap dance? *ERS Spectrum, 27*(3), 31-40.

Heneman, H. G., III, Milanowski, A., Kimball, S. M., & Odden, A. (2006).*Standards-Based Teacher Evaluation as a foundation for knowledge- and skill-based pay* (pp. 1-16, Issue brief No. RB-45). Philadelphia, PA: Consortium for Policy Research in Education. Retrieved from http://eric.ed.gov/?id= ED493116 (ERIC Document Reproduction Service No. ED493116)

Hull, J. (2011) *Building a better evaluation system: At a glance*. Center for Public Education. Retrieved from http://www.centerforpubliceducation.org/Main-Menu/Evaluating-performance/Building-A-Better-Evaluation-System

Jehlen, A. (2011). *NCLB: Is it working?* National Education Association. Retrieved from http://www.metproject.org/downloads/MET_Gathering_ Feedback_ Research_Paper.pdf

Johanningmeier, E.V. & Richardson, T. (2008). *Educational Research, The National Agenda, and Educational Reform, A History*. United States: Information Age Publishing.

Johnson, B., & Christensen, L. B. (2008). *Educational research: Quantitative, qualitative, and mixed approaches* (Third ed., p. 604). Los Angeles: Sage Publications.

Kane, Rockoff, & Staiger (2006). *What does certification tell us about teacher effectiveness?* [working paper]. Retrieved from http://tntp.org/assets/ documents/TNTP_FactSheet_TeacherExperience_2012.pdf

Kane, T.L. & Staiger, D.O. (2010). *Gathering feedback for teaching* [research report]. Retrieved from http://www.metproject.org/downloads/MET_Gathering _Feedback_Research_Paper.pdf

Keesler, V., & Howe, C. (2012). *Understanding educator evaluations in Michigan* (Rep.). Michigan Department of Education. Retrieved from https: //www.michigan.gov/documents/mde/Educator_ Effectiveness_Ratings_Policy_Brief_403184_7.pdf

Kohn, A. (2000). *The case against standardized tresting: Raising the scores, ruining the schools*. Portsmouth, NH: Heinemann.

Krajewski, R.J., & Anderson, R.H. (1980, May). Clinical supervision a decade later. Association for Supervision and *Curriculum Development* 420-423.

Laitsch, D. (2005). A policymaker's primer on testing and assessment. (Assessment Policy, Publication No. 42). Alexandria, VA: ASCD.

Lunenburg, F. C., & Irby, B. J. (2008). *Writing a successful thesis or dissertation: Tips and strategies for students in the social and behavioral sciences*. Thousand Oaks, CA: Corwin Press.

Marshall, K. (2012) Fine-Tuning Teacher Evaluations. *Educational Leadership* 70.3 (2012): 50-53. *ASCD*. Web. 15 Mar. 2014.

Marzano, R.J. (2011). Examining the role of teacher evaluation in student achievement (White paper). Marzano Center for Teacher and Leadership Evaluation. Retrieved from http://www.oregoned.org/images/ pages/Marzano_White_Paper_on_role_of_Teacher_Evaluation_in_Student_Achievement.pdf

Marzano, R.J. (2012). The two purposes of teacher evaluation. *Educational Leadership*, *70*(3), 14-19. Retrieved from http://www.ascd.org/ publications/educational-leadership/nov12/vol70/num03/The-Two-Purposes-of-Teacher-Evaluation.aspx

Marzano, R.J. (2012, May). Teacher Evaluation Model. Retrieved from tpep-wa.org/wp-content/uploads/Marzano_Teacher_Evaluation_Model.pdf

Matzat, A.L. (n.d.) Massachusetts education laws. Retrieved from https://www3.nd.edu/~rbarger/www7/masslaws.html

Medley, D. & Coker, H. (1987). The accuracy of principals' judgments of teacher performance. *Journal of Educational Research*, 80(4). doi:10.1080 /00220671.1987.10885759

Michigan Council for Educational Effectiveness, Building an Improvement focused System of Educator Evaluation in Michigan. (2013). Retrieved from http://www.mcede.org midnightreport_july24_2013-4.pdf

Michigan Department of Education, Office of Assessment & Accountability (2004-2005). *MEAP Coordinator Handbook* Lansing: Michigan Department of Education. Retrieved from http://www.michigan.gov/documents/F04 _OEAA_Coord_Manual_105606_7.pdf

Michigan Department of Education (2007). *Grade 3 released items* Lansing: Michigan Department of Education.

Michigan Department of Education. (2007). *Highly qualified teacher*. Retrieved from http://www.michigan.gov/documents/mde/2007_NCLB_Highly_Qualified_Teacher_Update_197620_7.pdf

Michigan Department of Education, Michigan K-12 Standards Mathematics (2010). *Report of the Michigan K-12 Standards*. Retrieved from http://www.michigan.gov/documents/mde/K-12_MI_Math_Standards_REV_470033_7.pdf

Michigan Department of Education, Bureau of Assessment and Accountability & Measurement Incorporated (2011-2012). *MEAP Technical Report*. Lansing: Retrieved from http://www.michigan.gov/documents/mde/MEAP_2010-2011_Technical_Report_394693_7.pdf

Michigan Department of Education. (2014) *Educator evaluations and effectiveness in Michigan*. [Data set analysis]. Retrieved from http://www.michigan.gov /documents/mde/2013-14_Educator_Evaluations_ and_Effectiveness _485909_7.pdf

Michigan Department of Education. (2014) *Michigan schools accountability and scorecard*. https://www.michigan.gov/documents/mde/ScorecardGuide _426897_7.pdf

Michigan Legislature. (2014) Section 380.1249. Revised school code act 451 of 1976. Section 1249 2-e iii). Retrieved from http://www.legislature.mi.gov/(S (mae5bpecr5vjntk3drwgwrpd)/mileg.aspx?page=GetObject&objectname=mcl-380-1249

Michigan Legislature. (2011) Public Act 102 of 2011. Section 1248. Retrieved from http://www.legislature.mi.gov/%28S%28cxpg4tkildcfkqokh2t4waef%29%29/mileg.aspx?page=getobject&objectname=mcl-380-1248

Milanowski, A.T. (2011). *Validity on teacher evaluation systems based on the framework for teaching*. (working paper). University of Wisconsin-Madison. Retrieved from http://files.eric.ed.gov/fulltext/ED520519.pdf

Milanowski, A.T., Kimball, S.M., White, B. (2004). *The relationship between standards-based teacher evaluation scores and student achievement: replication and extensions at three sites.* (Working paper). University of Wisconsin. Retrieved from http://cpre.wceruw.org/papers/3site_long_te_sa_aera04te.pdf

Mooney, J. School Districts Comparison Shop for Teacher Evaluation Systems. *NJ Spotlight Newsletter* (17 Oct. 2012). Print.

Murchison, Carl. (Ed.) (1930). *History of Psychology in Autobiography* (Vol. 2, pp. 381-407). Republished by the permission of Clark University Press, Worcester, MA.

National Board for Professional Teaching Standards. (2014). History of Standards. *The beginnings of a movement*. Retrieved from http://www.nbpts.org/history

National Board for Professional Teaching Standards. (March 19, 2015). Two new studies add to the evidence base on board certified teachers' impact on student achievement [online forum]. Retrieved from http://www.nbpts.org/newsroom/two-new-studies-add-evidence-base-board-certified-teachers-impact-student-achievement

National Commission on Excellence in Education 1983 *A Nation at Risk: The Imperative for Educational Reform.* Washington, D.C. : U.S. Government Printing Office. Retrieved from http://www2.ed.gov/pubs/NatAtRisk/index.html

National Council on Teacher Quality. (2014, October). *Teacher evaluation policy in Michigan: Where is Michigan in implementing teacher effectiveness policies?* (Issue brief). Washington D.C. National Council on Teacher Quality. Retrieved from http://www.nctq.org/dmsView/Evaluation _Timeline_Brief_Michigan

National Education Association (2010) *Teacher Assessment and Evaluation*. [White paper] Retrieved May 5, 2015 from http://www.nea.org/assets/docs/HE/ TeachrAssmntWhtPaperTransform10_2.pdf

National Education Association (2010). Beyond two tests scores: Multiple measures for student learning and school accountability. [Policy brief] Retrieved from http://www.nea.org/assets/docs/PB38beyondtwotestscores2011.pdf

National Education Association. (2011). *Promoting and Implementing the National Education Association Policy Statement on Teacher Evaluation and Accountability* [Press release]. Retrieved April 11, 2015, from http://www.nea.org/assets/docs/2011NEA_Teacher_Eval_Toolkit.pdf

New Teacher Project. (2010, October). *Teacher Evaluation 2.0*. (Issue brief). Retrieved from http://tntp.org/assets/documents/Teacher-Evaluation-Oct10F.pdf

Nolan, J.F. & Hoover, L.A. (2004). *Teacher supervision*. New York: John Wiley and Sons/Jossey-Bass.

Nowlin, M. (2011, September). Michigan council for educator effectiveness [Web log post]. Retrieved April 10, 2015, from http://www.mcede.org/MCEE_Interim ProgressReport_Apr2012.pdf

Office of Assessment and Accountability MEAP, Educational Assessment and Accountability. (2005). *MEAP coordinator handbook* (pp. A1-H10). Retrieval from Michigan Department of Education http://www.michigan.gov/documents /F04_OEAA_Coord_Manual_105606_7.pdf

Palardy, G.J., & Rumberger, R.W. (2008). Teacher effectiveness in first grade: The importance of background qualifications, attitudes, and instructional practices for student learning. *Educational Evaluation and Policy Analysis, 30*(2), 111-140.

Perrone, V. (1991). On standardized testing. A position paper of the Association for Childhood Education International. *Childhood Education, 67*, 131-142. doi: 10.1080/00094056.1991.10521597

Phelps, R. (Ed.). (2004). *Defending Standardized Testing*. Mahwah, New Jersey: Lawrence Erlbaum Publishers.

Popham, J.W. (1999, March). Why standardized tests don't measure educational quality. *Educational Leadership*, 56(6), 8-15.

Porter, E. (2015, March 24). Grading teachers by the test. *New York Times*. Retrieved from http://www.nytimes.com/2015/03/25/business/economy/grading-teachers-by-the-test.html?_r=0

Radunzel, J., & Noble, J. (2012). *Predicting college grades from ACT assessment scores and high school course work and grade information* (pp. 1-88, Rep. No. ED542027). (ERIC Document Reproduction Service).

Reardon, S. F., Robinson-Cimpian, J. P., & Weathers, E. S. (2008) Patterns and trends in racial/ethnic and socioeconomic academic achievement gaps. *Handbook of research in education finance and policy*, 497-516.

Reddell, S. (2010, November 20). *High Stakes Testing: Our Children at Risk* [Scholarly project]. In *Eric.ed.gov*. Retrieved April 8, 2014, from eric.ed/gov/?q=high+stakes+testing

Riley, R. W. (1995). The Improving America's Schools Act of 1994 Reauthorization of the Elementary and Secondary Education Act. Washington, D. C., U. S. Department of Education. Retrieved from https://www2.ed.gov/offices/OESE/archives/legislation/ESEA/brochure/iasa-bro.html

Roeber, E. (2009). *Using tests to evaluate classroom teachers*. Michigan State University. Retrieved from http://www.michiganassessmentconsortium.org/sites/default/files/MAC-Whitepaper-Roeber-Classroom-Evaluation.pdf.

Roberts, C. M. (2004). *The dissertation journey*. Thousand Oaks, CA: Corwin Press.

Roberts, C. M. (2010). *The dissertation journey*. Thousand Oaks, CA: Corwin Press.

Robinson M. J. V., Lloyd, C. A., Rowe, K. J. (2008, December). The impact of leadership on student outcomes: an analysis of the differential effects of leadership types. *Educational Administration Quarterly Vol. 44, No. 5 (December 2008) 635-674* retrieved from http://www.palmbeachschools.org/dre/documents/meta_lead_2008.pdf

Rockoff, J. E., & Speroni, C. (2011). Labour Economics. *Subjective and Objective Evaluations of Teacher Effectiveness: Evidence from New York City*, 687-696. Retrieved from http://blogs.edweek.org/edweek/teacherbeat/10%2017%20rockoff_speroni_labour_econ_published.pdf

Rothman, R. & Barth, P. (2009, November). Center for Public Education. *Does Highly Qualified Mean Highly Effective?* Retrieved from http://www.centerforpubliceducation.org/Main-Menu/Staffingstudents/How-good-are-your-teachers-Trying-to-define-teacher-quality/Does-highly-qualified-mean-highly-effective.html

Salkind, N.J. (2008). *Encyclopedia of educational psychology*. doi**:** 10.4135/9781412963848

Sartain, L., Stoelinga, S. R., & Brown, E. R. (2011). Rethinking teacher evaluation in Chicago: Lessons learned from classroom observations, principal-teacher conferences, and district implementation (Research Report). Retrieved from The University of Chicago Consortium on Chicago School Research website: http://www.joycefdn.org/assets/1/7/Teacher-Eval-Report-FINAL1.pdf

Stickler, L. (2007). *A critical review of the SAT: Menace or mild-mannered measure?* The College of NewJersey journal of student scholarship. Vol. IX, (April 2007). Retrieved from http://joss.pages.tcnj.edu/files/2012/04/2007-Stickler-SAT-Critical-Review.pdf

Silver, Strong, & Associates, The Thoughtful Classroom Teacher Effective Framework. (2014) Retrieved from Retrieved from http://www.thoughtful classroom.com/PDFs/TCTEF/TCTEF_Basic_Rubric.pdfhttp:// www.edweek .org/ew/articles/2013/01/08/17teach_ep.h32.html?tkn=LLMFMHoLB

Sokal, M. M. (1987). *Psychological testing and American Society: 1890-1930: Symposium: 150th National meeting: Papers. New Brunswick,* NJ: Rutgers University Press.

State of the States. (2011). *Trends and early lessons on teacher evaluation and effectiveness policies*. Washington, DC: National Council on Teacher Quality. Retrieved from http://www.edweek.org/media/nctq_stateof thestates_embargoed.pdf

State of the States (2013). Connect the dots: *Using evaluations of teacher effectiveness to inform policy and practice*. Washington, DC: National Council on Teacher Quality. Retrieved from http://www.nctq.org/dmsView/State_of_ the_States_2013_Using_Teacher_Evaluations_NCTQ_Report

Stufflebeam, D.L. (1998). *The Personnel Evaluation Standards*. Newbury Park: CA. Sage Publications.

Stull Bill of 1971, AB 293, California Education Codes No. 44600-44665, (Stull, 1971).

Supovitz, J.A., & Poglinco, S. M. (2001). Instructional Leadership in a Standards-based Reform.

Supovitz, J.A. (2015). Is high-stakes testing working? *Penn Graduate School of Education*. Retrieved from https://www.gse.upenn.edu/review/feature/supovitz

Toch, T., & Rothman, R. (2008). *Rush to judgment: Teacher evaluation in public education*. Washington, DC: Education Sector. Retrieved from http://www.educationsector.org/publications/rush-judgment-teacher-evaluation-public-education.

Toppo, G. (2013, October 17). Study: Program in D.C. removed bad teachers. Retrieved from http://www.usatoday.com/story/news/nation /2013/10/17/ impact-washington-dc-shools/2993233/

Tucker, P. D., & Stronge, J. H. (2005). *Linking teacher evaluation and student learning*. Alexandria, VA: Association for Supervision and Curriculum Development.

United States Department of Education. (2002). *The No Child Left Behind Act of 2001*. Retrieved from http://www2.ed.gov/policy/elsec/leg/esea02/index.html.

United States Department of Education. (2004). *New No Child Left Behind Flexibility: Highly Qualified Teachers*. Retrieved from http://www2.ed.gov/nclb/methods/teacher/hqtflexibility.html.

United States Department of Education. (2007). A Nation at accountable twenty-five years after a Nation at Risk. Retrieved https://www2.ed.gov/rschstat/research/pubs/accountable/accountable.pdf

United States Department of Education (2009). Race to the Top program executive summary. Washington, DC: Author. Retrieved from www2.ed.gov/programs/racetothetop/executive-summary. Pdf

United States Department of Education (2011). *Elementary and Secondary Education Act Flexibility Request*. Washington D.C. Retrieved from USED_esea-flexibility-request_372803_7.doc

United States Department of Education (2013). *Promoting evaluation rating accuracy strategic options for states*. Washington D.C: Reform Support Network.

Virginia Department of Education (n.d). *Connecting teacher evaluation to student achievement* 7-44. (Brief) Retrieved from http://va-sig-training. wmwikis. net/file/view/Briefs.pdf

Weiss E.M., & Weiss, S. (1998). New directions in teacher evaluation. Retrieved from http://www.ericdigests.org/1999-4/new.htm (ED429052).

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness*. Brooklyn, NY: New Teacher Project. Retrieved from http://widgeteffect.org/downloads/The WidgetEffect.pdf

Wigdor, A. K., & Green, B. F. (1991). *Performance assessment for the workplace* (Vol. 2). Washington, D.C.: National Academy Press.

Zimco. (2012). *K-12 Evaluation Solutions* [Brochure]. Author. Retrieved February, 2015, from http://www.k12evaluationsolutions.com/stages-about/company-background

Appendices

Appendix A
District Internal Research Application Request

District Internal Research Application Request

Request to Use [redacted] District Data

Thank you for your interest in conducting an external study involving the [redacted] District. Requests are reviewed by the Superintendent of Schools or his designee, and by the Department of Curriculum and Instruction. Research studies include surveys of students and staff, observations conducted in schools, analyses of existing school or student data, and pilot testing assessment instruments or rating scales. Individuals seeking to conduct an external research study using data from the [redacted] District must submit the following:

- • One completed copy of the application form
  Emailed and attached to: [redacted] Administrative Assistant to the Superintendent

Applicants are notified of the decision as soon as it has been reviewed.

Approval by the [redacted] District is contingent upon approval of directors, principals, teachers, students, and completion of informed consent forms by parents, as appropriate. The approval of the study does not constitute an endorsement of the study, and such language should not be included in the final reports.

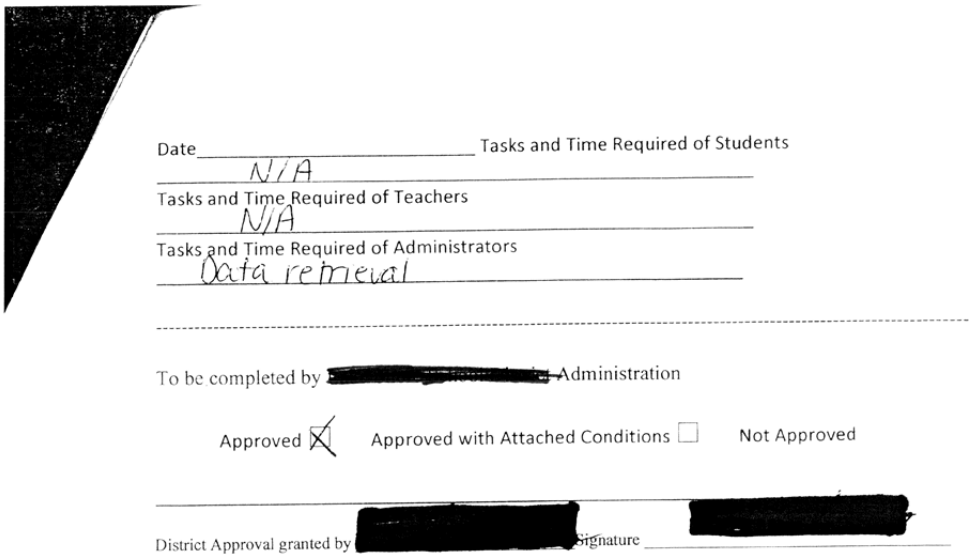
Any statistical reports must display the following disclaimer: "Statistics reported were prepared especially for this study and may not agree with other published statistics." All individuals who serve as members of the research team (e.g., applicant assistants, collaborators) and are not currently employed by the [redacted] District but may have unsupervised contact with students must complete the fingerprinting and background check required by the district/state.

The disruption of the school's routine by the study must be kept to a minimum and avoid any day in which standardized or district tests are administered.
Permission for research studies is for one year unless otherwise noted in the approval letter. Data collected is to be used solely for the purpose stated in the research application.

OUTSIDE RESEARCH EVALUATION APPLICATION AND APPROVAL FORM

Date: 10-9-2014
Applicant Name: Tammy DiPonio
Email: tdiponio@gmail.com Telephone: 816-225-5258 Cell Phone: 248-823-3728 Address: 4547 Quarton
City Bloomfield Hills State MI Zip 48301 Sponsor/Committee Chair Dr. Sharon Zoellner
University/Organization Research/Project Baker University
✓ Purpose of the research (reason for conducting research e.g. (dissertation), grant, & etc.)
Date of Actual Research: Starting Date July 2015 Completion

District Internal Research Application Request



Date _____ Tasks and Time Required of Students

N/A

Tasks and Time Required of Teachers

N/A

Tasks and Time Required of Administrators

Data retrieval

To be completed by [redacted] Administration

Approved ☒ Approved with Attached Conditions ☐ Not Approved

District Approval granted by [redacted] Signature [redacted]

District Internal Research Application Request

[redacted]

10-8-2014

Dear Institutional Review Board:

The purpose of this letter is to inform you that I give *Tammy DiPonio* permission to conduct the research using data from the [redacted] This also serves as assurance that this school complies with requirements of the Family Educational Rights and Privacy Act (FERPA) and the Protection of Pupil Rights Amendment (PPRA) (see next page for specific requirements) and will ensure that these requirements are followed in the conduct of this research.

Sincerely,

[redacted]

Superintendent, [redacted]

District Internal Research Application Request

[redacted]

Protection of Pupil Rights Amendment (PPRA):

- The right of a parent of a student to inspect, upon the request of the parent, a survey created by a third party before the survey is administered or distributed by a school to a student.
- Any applicable procedures for granting a request by a parent for reasonable access to such survey within a reasonable period of time after the request is received.
- Arrangements to protect student privacy that are provided by the agency in the event of the administration or distribution of a survey to a student containing one or more of the following items (including the right of a parent of a student to inspect, upon the request of the parent, any survey containing one or more of such items): Political affiliations or beliefs of the student or the student's parent. Mental or psychological problems of the student or the student's family. Sex behavior or attitudes. Illegal, anti-social, self-incriminating, or demeaning behavior. Critical appraisals of other individuals with whom respondents have close family relationships. Legally recognized privileged or analogous relationships, such as those of lawyers, physicians, and ministers. Religious practices, affiliations, or beliefs of the student or the student's parent. Income (other than that required by law to determine eligibility for participation in a program or for receiving financial assistance under such program).
- The right of a parent of a student to inspect, upon the request of the parent, any instructional material used as part of the educational curriculum for the student. Any applicable procedures for granting a request by a parent for reasonable access to instructional material received.
- The administration of physical examinations or screenings that the school or agency may administer to a student.
- The collection, disclosure, or use of personal information collected from students for the purpose of marketing or for selling that information (or otherwise providing that information to others for that purpose), including arrangements to protect student privacy that are provided by the agency in the event of such collection, disclosure, or use.
- Names of any sort for students and staff are to be eliminated in the publication of the study and should be discretely coded by the data processing department before given to researcher.
- The right of a parent of a student to inspect, upon the request of the parent, any instrument used in the collection of personal information before the instrument is administered or distributed to a student.

Appendix B: Baker University IRB Proposal for Research Permission Form and Approval

Baker University IRB Proposal for Research Permission Form and Approval

BAKER UNIVERSITY

SCHOOL OF EDUCATION
GRADUATE DEPARTMENT

Date: May 15, 2015
IRB PROTOCOL NUMBER _____
(IRB USE ONLY)

IRB REQUEST
Proposal for Research
Submitted to the Baker University Institutional Review Board

I. Research Investigator(s) (Students must list faculty sponsor first)

Department(s) **School of Education Graduate Department**

	Name	Signature	
1.	Sharon Zoellner, PhD.		Major Advisor
2.	Katie Hole, PhD.		Research Analyst
3.			University Committee Member
4.			External Committee Member

Principal Investigator:
Tammy DiPonio
Phone: 816-225-5258
Email: tdiponio@gmail.com
Mailing address: 5457 Quarton Rd. Bloomfield Hills, MI 48301

Faculty sponsor: Dr. Sharon Zoellner
Phone: 913-344-1225
Email: Sharon.Zoellner@Bakeru.edu
Expected Category of Review: ____Exempt _X_Expedited ___ Full

II: Protocol Title

Relationship Between Teacher Evaluation Scores and Student Achievement

Summary

In a sentence or two, please describe the background and purpose of the research.

Each year in Michigan certified teachers are evaluated and receive a score in one of four areas; highly effective, effective, minimally effective, and ineffective. Across the state less than 2% of teachers were rated *minimally effective* or *ineffective* in 2011-2012. The other teacher evaluation scores fell into the top two categories, *highly effective* and *effective*. The first purpose of the study is to determine the relationship between student performance on standardized tests and teachers who were rated *highly effective, effective, minimally effective,* and *ineffective*. The second purpose was to determine if there is a relationship between teacher experience and evaluation ratings and if these could be used to predict student performance as measured by the Michigan Education Assessment Program (MEAP).

Briefly describe each condition or manipulation to be included within the study.

No conditions or manipulations were done in this study. Conclusions will be made based on archived data.

What measures or observations will be taken in the study? If any questionnaire or other instruments are used, provide a brief description and attach a copy.

MEAP data and teacher evaluation ratings will be used in the study. No questionnaires or other instruments will be used in the study.

Will the subjects encounter the risk of psychological, social, physical, or legal risk? If so, please describe the nature of the risk and any measures designed to mitigate that risk.

No psychological, social, physical, or legal risks will be involved in this study.

Will any stress to subjects be involved? If so, please describe.

Participants will not be used in the study therefore there will be no stress involved.

Will the subjects be deceived or misled in any way? If so, include an outline or script of the debriefing.

Subjects will not be deceived or misled in any way.

Will there be a request for information that subjects might consider to be personal or sensitive? If so, please include a description.

There will not be a request for information which might be considered to be personal or sensitive.

Will the subjects be presented with materials that might be considered to be offensive, threatening, or degrading? If so, please describe.

There will not be subjects presented with materials that might be considered to be offensive, threatening, or degrading.

Approximately how much time will be demanded of each subject?

The study will not involve any subjects, only data; therefore there will not be any time involved for subjects.

Who will be the subjects in this study? How will they be solicited or contacted? Provide an outline or script of the information which will be provided to subjects prior to their volunteering to participate. Include a copy of any written solicitation as well as an outline of any oral solicitation.

Data will be used in the study instead of subjects therefore it is not necessary to contact any subjects.

What steps will be taken to ensure that each subject's participation is voluntary? What if any inducements will be offered to the subjects for their participation?

No subjects will be contacted for this study therefore it is not necessary to offer voluntary participation. All data for the study will be coded anonymously. There will be no inducements offered for participation.

How will you ensure that the subjects give their consent prior to participating? Will a written consent form be used? If so, include the form. If not, explain why not.

Written consent is not necessary, as no subjects will be contacted for this study. All data for the study will be coded anonymously.

Will any aspect of the data be made a part of any permanent record that can be identified with the subject? If so, please explain the necessity.

The data analyzed is already part of the district's permanent records. Results will not be included in the permanent records.

Will the fact that a subject did or did not participate in a specific experiment or study be made part of any permanent record available to a supervisor, teacher or employer? If so, explain.

No permanent records indicating that data was used will be available to the supervisors, teachers, or employer.

What steps will be taken to ensure the confidentiality of the data?

The researcher will not have access to any names; all data will be reported and coded by the district data analyst to maintain anonymity. According to Baker guidelines all data are kept in a locked or password protected computer. Three years after the study the data is destroyed.

If there are any risks involved in the study, are there any offsetting benefits that might accrue to either the subjects or society?

No risks or benefits are involved in the study. The district may benefit from the findings through analyzing the relationship between standardized tests results and teacher performance ratings. Utilizing this data with other instructional data may provide evidence for instructional programming needs and validation of programs already in place. Conclusions from class makeup in relation to student performance results, teacher years of experience, and teacher performance ratings may also be drawn from this study.

Will any data from files or archival data be used? If so, please describe.

Yes, data from the teacher evaluation ratings, student MEAP scores for math and language arts, and teachers' years of experience will be retrieved from the districts' data warehouses. Data used in the study is archival and is available to district employees with access and authorization to the data. For the purpose of this study, the data will be generated from the district data warehouse for each of the cohort years in the study and for each of the studies' categories. All data will remain anonymous.

BAKER UNIVERSITY
Own Confidence

Baker University Institutional Review Board

May 22, 2015

Dear Tammy DiPonio,

The Baker University IRB has reviewed your research project application and approved this project under Exempt Status Review. As described, the project complies with all the requirements and policies established by the University for protection of human subjects in research. Unless renewed, approval lapses one year after approval date.

Please be aware of the following:

1. Any significant change in the research protocol as described should be reviewed by this Committee prior to altering the project.
2. Notify the IRB about any new investigators not named in original application.
3. When signed consent documents are required, the primary investigator must retain the signed consent documents of the research activity.
4. If this is a funded project, keep a copy of this approval letter with your proposal/grant file.
5. If the results of the research are used to prepare papers for publication or oral presentation at professional conferences, manuscripts or abstracts are requested for IRB as part of the project record.

Please inform this Committee or myself when this project is terminated or completed. As noted above, you must also provide IRB with an annual status report and receive approval for maintaining your status. If you have any questions, please contact me at CTodden@BakerU.edu or 785.594.8440.

Sincerely,

Chris Todden EdD
Chair, Baker University IRB

Baker University IRB Committee
Verneda Edwards EdD
Sara Crump PhD
Erin Morris PhD
Scott Crenshaw