THE DIFFERENCES IN STUDENT ACHIEVEMENT BASED ON MISSOURI APPROVED TEACHER EVALUATION MODELS USED

TERESA M. ADAMS

2020

The undersigned, approved by the Department Chair of Graduate Studies in Education, have examined a dissertation entitled:

THE DIFFERENCE IN STUDENT ACHIEVEMENT BASED ON TEACHER EVALUATION MODEL USED TO EVALUATE TEACHERS IN THE STATE OF MISSOURI

Presented by, Teresa M. Adams, a candidate for the degree of Doctor of Education and hereby certify that in their opinion it is worthy of acceptance.

Dr. Allison Langford, Advisor/Chair

Graduate Education, Southwest Baptist University

Dr. Tammy Condren, Committee Member

Graduate Education, Southwest Baptist University

Dr. Benny Fong, Committee Member

Assistant Professor of Graduate Education and Statistics, Southwest Baptist University

THE DIFFERENCES IN STUDENT ACHIEVEMENT BASED ON MISSOURI APPROVED TEACHER EVALUATION MODELS USED

A Dissertation
Presented to
The Faculty of the Graduate Education Department
Southwest Baptist University

In Partial Fulfillment of the
Requirements for the Degree
Doctor of Education

By

Teresa M. Adams, B.S., M.S., Ed. Spec.

Dr. Allison Langford, Dissertation Advisor

May 16, 2020

ACKNOWLEDGEMENTS

Completing this study could not have been accomplished without the help, support, and encouragement of many incredibly amazing people. It was nothing short of a blessing to have Dr. Allison Langford advise me during this process. Dr. Langford always knew what I needed and when I needed it. Her kindness, patience, and encouragement kept me going when my confidence waned. Her words were always positive, even when she had to be direct. I aspire to be the kind of advisor to others that she has been to me and thank her for all she has done. I also owe a debt of gratitude to Dr. Benny Fong and Dr. Tammy Condren for their support through this process. Dr. Fong challenged me to stretch my thinking beyond what I ever imagined I could. If not for him, I would never have made it through Chapter Four. Dr. Condren's feedback throughout my paper kept me on track, encouraging me to expand my thinking and be thorough and current in my research. I thank them both.

Throughout this time working on my paper, my family has been a constant source of support. No one believes in me more! I thank my mother, my first, best, and most steadfast supporter. She believes in me, knows me better than I know myself, and understands that completing this paper was about much more than the degree. I also thank Charles, for encouraging me, even when he wasn't well, and having the best paper writing recliner ever built. My sister, Cathi, a fellow elementary principal, is my built-in sounding board and always up for a philosophical discussion. My brother, Dave, always makes me laugh, just when I need to and finds ways to let me know he's proud of me. For my niece, Madison, and my nephews, William, Joseph, Joshua, and Jackson, you can accomplish anything you set your mind to, and I will be your biggest supporter!

Finally, I would like to thank three people without whom I would never have decided to take this journey in the first place. We barely knew each other when we started, became a team of colleagues, and now I count them as my closest friends. I miss those Saturday car rides! If not for Amanda Boyer, Jason Weaver, and Jill White, the "exciting journey from ordinary citizen to doctor" (Truelove, 2014) would have been much less entertaining.

TABLE OF CONTENTS

LIST OF TABLES

ABSTRACT

In response to legislation from the federal government to consider student achievement when evaluating teachers, states have created new evaluation tools. The challenge is in isolating the many variables, including the actions of the teacher, that may impact student achievement. The purpose of the causal-comparative study was to determine the difference in the percent of students scoring proficient and advanced on the Missouri Assessment Program (MAP) in third, fourth, and fifth grades, depending on which evaluation model was used to evaluate teachers between the years 2017 and 2019. In response to No Child Left Behind, the Missouri Department of Elementary and Secondary Education (MODESE) required school districts to use the DESE created Model Educator Evaluation System (MEES), or use an evaluation system based on the same principles used to create the MEES and approved by DESE. This study compared the mean MAP scores of students in third, fourth, and fifth grades in 2017, 2018, and 2019 whose teachers were evaluated using either the MEES, the Network for Educator Effectiveness (NEE), or a district created evaluation, to determine the differences in the mean scores. School districts who used the same evaluation for each of the three years were selected for the study and their total percent of proficient and advanced scores were obtained from MODESE's open-access website. A one-way ANOVA was used to compare the means. The findings of the study showed statistically significant differences in 2017 fourth grade ELA, 2017 fourth grade math, 2017 fifth grade ELA, 2017 fifth grade math, and 2018 fourth grade ELA. While there were statistically significant difference, the effect size was low. The outcomes of the study did show the MEES had the lowest mean percent of proficient and advanced scores in all but two of the eighteen

comparisons, indicating the need for additional research studies regarding the impact of teacher evaluation on student achievement.

Chapter One

Introduction

In 1965, President Lyndon B. Johnson signed into law the Elementary and Secondary Education Act, sharing his belief that giving children the chance to receive an education must be "our first national goal" (*Every Student Succeeds Act*, 2017). In 1983, "A Nation at Risk" was likely the catalyst for the increased focus and attention on education. A report to the United States conducted by the National Commission on Excellence in Education, "A Nation at Risk" warned that America could be in serious jeopardy of losing status in the world if attention was not paid to issues identified in education (1983). Among the most crucial issues, the committee noted low teacher pay, inadequate education for teacher preparation, decreasing test scores, and an increase in the number of Americans who were illiterate (1983). In a comparison with other countries on 19 tests, the United States never scored in the top two, and scored last seven times ("A Nation at Risk," 1983). The Elementary and Secondary Education Act has gone through several reauthorizations, including The No Child Left Behind (NCLB) under George W. Bush in 2001, and the latest reauthorization, the Every Student Succeeds Act (ESSA) under Barack Obama (*Every Student Succeeds Act*, 2017.). A key component of the NCLB and ESSA reauthorizations was the requirement that states hold districts accountable for student learning, measuring progress through standardized achievement tests. (*Every Student Succeeds Act*, 2017).

Holding schools accountable for student achievement means ensuring students are receiving the best possible instruction from high quality, effective teachers. According to Strong, Gargani, and Hacifazlioglu (2011), NCLB and Obama's Race to the Top, which

promised school districts additional funding for improved student achievement, forced schools to take a critical look at the effectiveness and ability of teachers to provide the environment and education needed for students to become proficient, as defined by each individual state. Missouri responded in 2013 through a news release, which announced the approval of the new Model Educator Evaluation System (MODESE, 2013b). In the Overview of Essential Principles of Evaluation (2013a), DESE identified seven principles on which the evaluation system is based. By 2014, all districts in Missouri were required to implement the Model Educator Evaluation System (MEES), the Network for Educator Effectiveness (NEE), or a DESE approved, district created evaluation that met DESE's criteria. One of the principles DESE required as part of the educator evaluation was the use of student growth on annual standardized tests to measure the educator's performance.

Theoretical Framework

In their work, *A Conceptual Framework for Examining Teachers' Views of Teaching and Educational Policies,* Darling-Hammond and Wise (1981) identified the Rationalistic Theory as one based on science. The Rationalistic Theory is based on the desire to affect behavior by applying very specific actions (Darling-Hammond & Wise, 1981). By giving teachers specific goals and objectives related to curriculum, directing teachers in which strategies can be used to meet the goals and objectives, and then evaluating teachers on how well the goals and objectives were met, evaluators looked to rationalize the behavior of the teachers (Darling-Hammond & Wise, 1981). Darling-Hammond and Wise believed education relies on the Rationalistic Theory from the science discipline because education lacked a true theory of its own.

In science, research is conducted through experiments on objects with known attributes, leading the researcher to have the ability to make reasonable predictions about behavior. Because of this predictability, the researcher is able to rationalize reality (Darling-Hammond & Wise, 1981). When applied to evaluation, the Rationalistic Theory assumes all students will respond in the same way when the same instructional strategy is applied by the teacher, resulting in the same outcome (Darling-Hammond & Wise, 1981). In the Rationalistic Theory, it is believed that once goals and objectives are set by administrators, experimentation with the application of varying instructional strategies will lead to predictable patterns of behavior, as is the case with scientific experiments (Darling-Hammond & Wise, 1981). An administered assessment of the student would determine how well the teacher implemented the strategy, not taking into account any action on the part of the student (Darling-Hammond & Wise, 1981).

The Marzano Focused Teacher Evaluation Model is the theoretical framework applied in this study to teacher evaluation. First developed in 2010, Marzano developed his model with the intent of providing a research-based evaluation for the purpose of teacher growth, which led to student achievement (2011). The Marzano Focused Teacher Evaluation Model is a standards-based evaluation model that goes a step further than just providing evaluators with a tool to accurately judge the performance of teachers; it also offers a system of support for the teacher, and provides a method for the evaluator to offer support, in an effort to increase both the teacher's skill and the student's achievement (Carbaugh, Marzano, & Toth, 2017).

The Marzano Model has four domains with 23 elements, or behaviors, distributed among the domains. Teachers are evaluated on each of the 23 elements throughout the

year, however they are not evaluated on all 23 elements at each evaluation or observation (Carbaugh, et al., 2017). The four domains include: "Standards-Based Planning (3 elements), Standards-Based Instruction (10 elements), Conditions for Learning (7 elements), and Professional Responsibilities (3 elements)" (Carbaugh, et al., 2017, p. 9). The Marzano Model begins with a pre-observation conference, at which the teacher describes the lesson planned, standards to be taught, and strategies that will be used. The evaluator works with the teacher to ensure the plan meets the correct level of rigor and standards and strategies are appropriate. The evaluator then visits the classroom for the observation. During this observation, the evaluator determines the teacher's score based on student evidence of learning. Following the observation, the teacher and evaluator meet for a post-observation conference, at which the teacher's performance, score, and student evidence are discussed. The teacher may present student evidence if the evaluator was unable to observe any evidence while in the classroom. At this time, the teacher and evaluator will determine if any follow up is needed to increase the teacher's skill and knowledge, based on the score received (Carbaugh, et al., 2017). Overall, examining teacher evaluation and student achievement through these theoretical frameworks leads to the problem described in the following section.

Problem Statement

The purpose of teacher evaluation systems is two-fold, supervision and evaluation (Mette, et al., 2017; Range, 2013). Supervision is typically seen as a support function, providing feedback to help the teacher improve, while evaluation is seen as the formal practice of holding the teacher accountable to standards (Mette, et al., 2017). An effective teacher is essential to and has a strong impact on student achievement. In 2009, John

Hattie provided a comprehensive look at research studies conducted to determine what had the largest impact on a student's academic success. Of the top 20 influences on student achievement, four were attributed to the teacher and eight were attributed to the act of teaching, all having an effect size of between 0.61 and 0.90. Because research shows the teacher has a positive impact on student achievement, which can potentially lead to greater than a year's academic growth for a student, evaluators need effective tools to provide guidance and feedback that will help the teacher improve and improve student achievement. It is unclear whether the evaluation systems DESE has approved lead to greater student achievement. Despite many changes to evaluation systems across the United States in response to the Race to the Top initiative and NCLB's promised incentives, there was minimal change in the number of teachers considered ineffective (Dee & Wyckoff, 2017). Hattie (2009) likely provided the most comprehensive list of variables to improve learning. However each of the variables cannot be measured during an evaluation of the teacher. The fact that there are so many variables that can impact learning in different ways leads to the purpose of this study.

In 2012, Missouri received a waiver from the requirements of NCLB, which allowed the state the opportunity to implement its own system of accountability within the state's school districts, which focused on school improvement and student achievement (MODESE, 2015). One component Missouri promised as part of the waiver was the development of a more comprehensive teacher evaluation system, which would hold teachers more accountable for student achievement (MODESE, 2015). This study examines the evaluation systems used by school districts in the state of Missouri to determine if there is a difference in student achievement in grades three, four, and five on

the Missouri Assessment Program (MAP) based on the evaluation system used. Considering the impact of teachers on student achievement and Missouri's response in creating a teacher evaluation system to hold teachers accountable for student achievement, the following rationale was the basis for this research study.

Rationale for the Study

The purpose of this causal-comparative study was to determine the differences in the percent of students scoring proficient and advanced on the Missouri Assessment Program MAP in third, fourth, and fifth grades, depending on which evaluation system was used to evaluate the teacher between the years 2017 and 2019. The independent variable of the study is the evaluation system used by each of the schools. The choices are the Model Educator Evaluation System (MEES), the Network for Educator Effectiveness (NEE), or a district created evaluation system. The dependent variable of the study is the sum of the proficient and advanced scores on the MAP scores for the years 2017 through 2019.

The use of student growth data, or value-added measures (VAMs), is one of the principles on which teachers must be evaluated according to DESE's Overview of Essential Principles of Evaluation (2013a). If the teacher's job requires students to achieve at high levels on state standardized tests, then the tools used to evaluate the teachers must be effective. While teachers appreciate efforts to provide more comprehensive evaluations and the use of multiple measures to provide meaningful feedback on instructional effectiveness in the classroom, teachers also express frustration with the inconsistent training and messages received from evaluators (Donaldson, 2016). Many new teacher evaluations require teacher creation and use of student learning

objectives. However teachers report little training on how to create student learning objectives. Additionally, teachers receive inconsistent messages about what they should include about how to develop the objectives; some supervisors coach teachers to create objectives that are simple and generic, while others coach teachers to set objectives high expectations (Donaldson, 2016).

Morgaen Donaldson (2016) found that reforms such as the No Child Left Behind and the Race to the Top initiative, have led to changes, especially with teachers evaluating their own work. However, concerns have been raised as to whether the changes are focused on the right issues and whether changes will lead to improved student achievement. One concern regarding goal setting centered on appropriate content and whether teachers were putting effort into setting goals that were essential for student learning and achievement. Another concern focused on the use of student performance data from assessments as a part of the evaluation process. To improve their evaluation results, teachers may put more effort into students who may have a more positive impact on their evaluation results than they do in other students. One last concern was in regard to the pacing of instruction. The concern of pacing was the teacher may slow the pacing of instruction to ensure mastery, which would lead to the detriment of teaching the comprehensive curriculum (Donaldson, 2016).

The Gates Foundation's Measures of Effective Teaching (MET) Project (2010) found, while there is agreement that effective teaching is necessary, there is a distinct lack of agreement on how to measure effective teaching. The MET further criticized current evaluation systems as lacking in providing necessary information to help close the achievement gap. The MET also cited teacher evaluation's lack of useful feedback to

help teachers improve, as well as the inability to provide supervisors with the necessary information to determine a teachers' effectiveness, including strengths and weaknesses. The MET findings echo the observations of 2009's *The Widget Effect*. Weisberg, Sexton, Mulhern, and Keeling conducted a study of 15,000 teachers, 1,300 administrators, and 12 districts and found, despite the efforts of states to improve evaluation systems following NCLB, evaluators continued to rate teachers in the highest categories by 94-99 percent (2009).

Considering the lack of research related to the impact of teacher evaluation on student achievement, states must begin to consider how to continue to mandate the use of value added measures (VAMs) and student growth measures to affect change, not just compliance (Taylor & Tyler, 2012; Mantzicopoulos, Patrick, Strati, & Watson, 2018). Taylor and Tyler (2012) encouraged states to move beyond compliance with federal mandates and place meaningful effort into taking a more optimistic approach; researching what works in evaluations to motivate teachers to change practices that do not work and adopt more effective practices. Putting effort into evaluation design, VAMs, frameworks, and implementation are crucial if districts want to be successful (Papay, 2012). In light of these findings and continued concerns that evaluators frequently continue to rate teachers in the top categories, research questions were developed.

Research Questions

The following research questions have been developed to address the stated problem:

1. What are the differences in the **third grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the

Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system, and a DESE approved, district developed evaluation model in **2017**?

2. What are the differences in the **fourth grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Missouri Educator Evaluation System, the Network for Educator Effectiveness evaluation system, and a DESE approved, district developed evaluation model in **2017**?
3. What are the differences in the **fifth grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system, and a DESE approved, district developed evaluation model in **2017**?
4. What are the differences in the **third grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system, and a DESE approved, district developed evaluation model in **2018**?
5. What are the differences in the **fourth grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system, and a DESE approved, district developed evaluation model in **2018**?

6. What are the differences in the **fifth grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system, and a DESE approved, district developed evaluation model in **2018**?
7. What are the differences in the **third grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system, and a DESE approved, district developed evaluation model in **2019**?
8. What are the differences in the **fourth grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system, and a DESE approved, district developed evaluation model in **2019**?
9. What are the differences in the **fifth grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system, and a DESE approved, district developed evaluation model in **2019**?

Null Hypotheses

The following null hypotheses were investigated to answer the research questions:

$H_0$1. There are no statistically significant differences in the **third grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system, and a DESE approved, district developed evaluation model in **2017**.

$H_0$2. There are no statistically significant differences in the **fourth grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system, and a DESE approved, district developed evaluation model in **2017**.

$H_0$3. There are no statistically significant differences in the **fifth grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system, and a DESE approved, district developed evaluation model in **2017**.

$H_0$4. There are no statistically significant differences in the **third grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system, and a DESE approved, district developed evaluation model in **2018**.

$H_0$5. There are no statistically significant differences in the **fourth grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for

Educator Effectiveness evaluation system, and a DESE approved, district developed evaluation model in **2018**.

$H_0$6. There are no statistically significant differences in the **fifth grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system, and a DESE approved, district developed evaluation model in **2018**.

$H_0$7. There are no statistically significant differences in the **third grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system, and a DESE approved, district developed evaluation model in **2019**.

$H_0$8. There are no statistically significant differences in the **fourth grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system, and a DESE approved, district developed evaluation model in **2019**.

$H_0$9. There are no statistically significant differences in the **fifth grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system, and a DESE approved, district developed evaluation model in **2019**.

Definition of Key Terms

For the purpose of this study, the following key terms and definitions were used:

Missouri Assessment Program (MAP) - An annual assessment required by Missouri's Department of Elementary and Secondary Education to determine student acquisition of the Missouri Learning Standards (MODESE, 2014).

MAP Grade Level Assessment - The annual assessment required by Missouri's Department of Elementary and Secondary Education for students in third through eighth grades, to determine students acquisition of the Missouri Learning Standards (MODESE, 2014).

Below basic - Score on the MAP that indicates a minimal understanding of the Missouri Learning Standards (MODESE, 2015).

Basic - Score on the MAP that indicates partial understanding of the Missouri Learning Standards (MODESE, 2015).

Proficient - Score on the MAP that indicates adequate understanding of the Missouri Learning Standards (MODESE, 2015).

Advanced - Score on the MAP that indicates a thorough understanding of the Missouri Learning Standards (MODESE, 2015).

Teacher evaluation - Gathering information about a teacher, then using the information to determine the teacher's value (Darling-Hammond, 1990).

Elementary and Secondary Education Act (ESEA) - The 1965 legislation signed by President Lyndon B. Johnson that involved the federal government in providing funding to school districts to ensure equity in educational opportunities. The

most notable component being Title I funding for school districts serving students from low-income families (Every Student Succeeds Act, 2017).

No Child Left Behind (NCLB) - The 2001 reauthorization of ESEA, which placed accountability requirements on each state by the federal government (Every Student Succeeds Act, 2017).

Every Student Succeeds (ESSA) - The 2015 reauthorization of ESEA returning much control of educational decisions back to the states, while maintaining a focus on accountability and equal educational opportunities for students identified as disadvantaged (Every Student Succeeds Act, 2017).

Effect size - A scale used to determine influence on achievement. An effect size of .4 is considered the point at which a positive influence is present and, in education, equal to one year of growth (Hattie, 2009).

Value Added Measure (VAM) - A score, based on student achievement on a high stakes standardized test, used to predict the effectiveness of the teacher (Sandilos, Sims, Norwalk, & Reddy, 2019).

Limitations

The following are limitations of the research:

1. There are many variables that can impact student achievement; however all variables will not be investigated for the purpose of this study.
2. This study is limited to the elementary student achievement in grades three, four, and five only.
3. This study is limited to students attending public school districts in Missouri.
4. Private and parochial schools are not subject to Missouri DESE evaluation

requirements, therefore results from this study will not apply to schools in these categories.

5. Standardized assessments are subject to limitations in the areas of validity and reliability, which can question the results.
6. School district's strict adherence to the guidelines and requirements set forth by the evaluation method used.
7. The study can only determine a correlation between the evaluation tool used and student achievement and not causation.

Delimitations

The following are the delimitations of the research:

1. Student performance on the MAP test was collected using Missouri DESE's Web applications site from 2017, 2018, and 2019, following Missouri DESE's 2014 deadline to adopt one of the approved evaluation methods.
2. Public school districts in the state of Missouri.
3. Student achievement in grades three, four, and five on the MAP during the years 2017 through 2019.
4. All public-school districts using the same evaluation system for the years 2017, 2018, and 2019 were included in the study as opposed to conducting a random sampling of school districts.
5. Rationalistic Theory was applied to student achievement.
6. Marzano's Focused Evaluation Model was applied to teacher evaluation.

Assumptions

The following are the assumptions made in the course of the research:

1. Missouri's Department of Elementary and Secondary Education will continue to require districts to use the DESE approved models of evaluation.
2. Supervisors understand the teacher evaluation system used by their school.
3. If differences are found for the students in grades three through five used in this study, then the same would be true for all students in grades three through five in the state of Missouri.

Design Controls

This causal-comparative study explored the differences in the percent of students in grades three, four, and five scoring proficient and advanced on the MAP, based on the type of evaluation system used to evaluate the teachers in each district. School districts in Missouri, using the Missouri Educator Evaluation System, the Network for Educator Excellence and a district created, Missouri DESE approved system for 2017, 2018, and 2019 were identified. Data for grades three, four, and five in each district was then collected from Missouri DESE's Comprehensive Data System to determine the percent of students scoring proficient and advanced in each grade level. The scores were then analyzed using a one-way analysis of variance (ANOVA) to compare the means of each grade level for each year to determine if there were significant differences in the percent of students scoring proficient and advanced depending on which evaluation system was used to evaluate the teacher.

The researcher addressed each of the limitations as many variables can impact student achievement. The researcher chose to focus on the impact of teacher evaluation for the purpose of this study due to Hattie's identification of the teacher as having the most influence on student achievement (2009). The study was limited to the achievement

of students in grades three, four, and five as those are the grade levels at which Missouri has consistently assessed students with the MAP. The researcher limited the study to students attending public school districts in the state of Missouri, as those are the only students required to take the MAP. For this reason, it will not be possible to apply the results of this study to students attending private and parochial schools. Results of the study are limited by the reliability and validity issues of standardized testing, largely due to the vast number of standards in the state of Missouri, and the inability of any standardized test to be able to represent how well the teacher taught or the student learned each standard. The study was limited by the evaluator's adherence to the guidelines and requirements of the evaluation method used. The researcher chose school districts that had been using the evaluation method for the 2017, 2018, and 2019 school years. Because this is a correlational study, the researcher was unable to determine if the evaluation of the teacher caused the achievement of the student. However, due to research determining a teacher's influence on student achievement, the researcher can determine a correlation between the evaluation of the teacher and the achievement of the student.

The researcher identified delimitations for the study. Student performance data on the MAP was collected only for the years 2017, 2018, and 2019, following DESE's 2014 deadline to implement one of the approved evaluation methods. These parameters ensured school districts had evaluators trained to use the new evaluation method as well as experience implementing the evaluation method. Only public-school districts in the state of Missouri were chosen for the study. Private and parochial schools are not required to take the MAP and would not have qualifying scores. Additionally, school

districts outside the state of Missouri would not take the MAP. Student achievement in grades three, four, and five on the MAP were selected as the researcher was interested in student achievement at the elementary level. The researcher obtained MAP results for all Missouri school districts using one of the three evaluation methods for each of the three years included in the study, therefore a stratified, purposive sampling was applied to these subgroups of school districts to select the districts included in the study. The Rationalistic Theory was applied to student achievement and the Marzano Focused Evaluation Model was applied to teacher evaluation.

Summary

Since the publication of "A Nation at Risk" in 1983, there has been a consistent focus on improving student achievement. Later evaluation reforms, including NCLB and Race to the Top, required states to begin to make changes to education policies, curriculums and evaluations in an effort to increase the achievement of their students. For most, this meant taking a serious look at current evaluation systems. Missouri restructured the evaluation system to incorporate the use of student achievement as part of the evaluation tool. While research, especially Hattie's meta-analysis, shows the teacher to have one of the most influential impacts on student learning, there is little evidence demonstrating the strategies and techniques used by the teacher are what lead to higher achievement scores on standardized assessments.

For the purpose of this research study, the theoretical frameworks applied were the Rationalistic Theory and Robert Marzano's Focused Evaluation. The following research questions were addressed:

1. What are the differences in the **third grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system, and a DESE approved, district developed evaluation model in **2017**?
2. What are the differences in the **fourth grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Missouri Educator Evaluation System, the Network for Educator Effectiveness evaluation system, and a DESE approved, district developed evaluation model in **2017**?
3. What are the differences in the **fifth grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system, and a DESE approved, district developed evaluation model in **2017**?
4. What are the differences in the **third grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system, and a DESE approved, district developed evaluation model in **2018**?
5. What are the differences in the **fourth grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness

evaluation system, and a DESE approved, district developed evaluation model in **2018**?

6. What are the differences in the **fifth grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system, and a DESE approved, district developed evaluation model in **2018**?
7. What are the differences in the **third grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system, and a DESE approved, district developed evaluation model in **2019**?
8. What are the differences in the **fourth grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system, and a DESE approved, district developed evaluation model in **2019**?
9. What are the differences in the **fifth grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system, and a DESE approved, district developed evaluation model in **2019**?

The purpose of this causal-comparative study was to determine the differences in the percent of students scoring proficient and advanced on the MAP in third, fourth, and fifth grades, depending on which evaluation system was used to evaluate their teacher between the years 2017 and 2019.

Chapter Two provides a review of the literature regarding teacher evaluation. It begins with the theoretical frameworks used for the purpose of this study. The Rationalistic Theory was used in reference to student achievement and Robert Marzano's Focused Evaluation was used in reference to teacher evaluation. The theoretical frameworks are followed by a brief history of teacher evaluation and how evaluation of teachers developed throughout the evolution and change of education in the United States. A review of education reform, including the influence of NCLB, Race to the Top, and ESSA provides an introduction to Missouri's decision to develop the state's own evaluation and require the new evaluation, or another state approved evaluation, to be used in Missouri public school districts. Finally, student achievement of elementary students is reviewed, as well as the history and goal of the MAP.

Chapter Two

Review of Related Literature

Introduction

Although recent federal legislation has created a sense of urgency for states to improve student achievement, concern for the influence schools have on the achievement of students is not new. While originally called upon to research and report regarding the disparities in educational opportunities received by white students and black students, James Coleman made two other discoveries in his 1966 report "1) families are the most important influence on student achievement, and 2) school resources don't matter" in the achievement of students (Hanushek, 2016, p. 19). Fifty years later, little had changed. Hanushek (2016) reported student achievement had grown by just .3 standard deviations in reading and .2 standard deviations in math. Additionally, The Coleman Report has led to conversations about what influences student achievement and has linked research to policy decisions (Hanushek, 2016). With The publication of The Coleman Report, the field of education has expanded through educational research.

Chapter Two provides a review of the literature related to teacher evaluation and student achievement. The theoretical frameworks used in this study are described and a brief history of teacher evaluation in the United States is summarized. Teacher evaluation and government reforms leading to change are addressed and Missouri's current evaluation requirements are outlined. An overview of student achievement and the Missouri Assessment Program's goal of measuring student achievement in Missouri elementary school districts is presented. The goal of this literature review is to provide

background and context for the study to determine if the teacher evaluation methods used in Missouri impact student achievement.

Theoretical Frameworks

The researcher explored the differences in student achievement on the MAP among school districts using the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system, and a DESE approved evaluation model. For the purpose of this study, two theories were applied, as there was not a single theory that addressed both student achievement and teacher evaluation. The Rationalistic Theory was applied to student achievement and the Marzano Focused Evaluation Model was applied to teacher evaluations.

Rationalistic Theory - Student achievement. Bhola (1990) explained the Rationalistic Theory as logical and positive. He also identified three components of the Rationalistic Theory. The first component, reductionism, occurs when part of an item can be taken from a whole without changing either object. The goal in reductionism focuses on the study of the individual components in an effort to better understand how the components work together as a whole. The second component, repeatability, occurs when an action performed by one person can be replicated by another. The final component, refutation, takes place when a conclusion can be either confirmed or refuted by another person (Bhola, 1990). Physical Science is the basis of the Rationalistic Theory, with the belief that predictions can be made based on the actions applied and the reactions observed over time can be repeated (Darling-Hammond & Wise, 1981).

When using the Rationalistic Theory in education, students are not perceived as active in the learning process, only as the receiver of the instruction the teacher delivers

to them. Provided the students have similar characteristics, such as age or grade placement, teachers should be able to predict the outcome of the strategies implemented to teach the material, based on previous results in similar situations with similar students (Darling-Hammond & Wise, 1981). When using the rationalistic approach, the goal is typically

> …(i) to make normative statements about populations based on randomly selected samples; (ii) to make comparisons between two groups, or before and after comparisons in regard to characteristics of the same group; and (iii) to establish correlations between characteristics of individuals or groups of individuals (Bhola, 1990, p. 210).

In a classroom setting, when the elements of Bhola's goal are met, predictions can be made regarding how students will perform on assessments, with the predictions based on past results which were repeated with similar students and confirmed in those results (Bhola, 1990; Darling-Hammond & Wise, 1981).

While the rationalistic approach continues to be applied in the field of education, it is considered by many to be too simple. A few of the simplifications of the Rationalistic Theory include: the belief that a uniform set of goals bring about the same results in all classrooms, all teachers will deliver instruction in the same way, and all students learn in the same way. The Rationalistic Theory allows no room for differences in teaching strategies or learning styles (Darling-Hammond & Wise, 1981).

Robert Marzano - Teacher evaluation. In the book *Effective Supervision*, Marzano, Frontier, and Livingston noted "the purpose of supervision should be the enhancement of teachers' pedagogical skills, with the ultimate goal of enhancing student

achievement" (2011, p. 2). Marzano believed evaluation systems should develop and grow teachers' skills rather than simply measure teaching abilities (2012). He emphasized evaluation systems should consist of three important features. First, the evaluation system should be comprehensive, containing features identified in research as proven to develop student achievement, and specific, including the strategies used and the behaviors of the teacher. Next, evaluations need to include a scale on which to score teachers so teachers may track progress toward developing specific skills. Finally, the system must include recognition and rewards for teacher growth toward the development and mastery of the skills (Marzano, 2012).

Marzano's development of his evaluation model came after years of research on instruction and teaching. Through his research, which Marzano wrote about in *What Works in Schools* (2005), *The Art and Science of Teaching* (2007), and *Classroom Instruction that Works*, written with Debra Pickering and Jane Pollock in 2001, Marzano identified strategies which when applied effectively by the teacher, yielded higher student achievement. Marzano pointed out that teachers could not just know which strategies worked, which he identified as the science of teaching, rather teachers also had to know the right time to use a particular strategy, which he identified as the art of teaching (2005). Using this research, Marzano created an evaluation model including the observation of strategies that are research based and proven to have a positive impact on student achievement when effectively implemented by the teacher (2011).

Marzano developed a framework for his observation model based not only on scoring not just teachers' demonstration specific strategies, but also incorporating student evidence and how the evidence supports the impact on student achievement. An equally

important part of Marzano's evaluation was providing feedback to the teacher following the observation for the purpose of teacher growth (2013). During evaluations, when evidence of strategies are not present or when teachers do not demonstrate proven strategies effectively, evaluators can provide specific feedback on necessary steps to improve. Previous models of teacher evaluation required long observations, more elements to observe, and much scripting, which were time consuming for the evaluator. Because Marzano's process was simplified, with only 23 elements and no required scripting, time gained by administrators using this model could potentially be spent following up with teachers on feedback given during the evaluation process, coaching teachers on specific instructional strategies, and providing professional development opportunities. More time to collaborate with teachers could potentially lead to improved instruction and greater student achievement (Marzano, 2012).

Over the years, Marzano's model has gone through several revisions and is known today as the Marzano Focused Teacher Evaluation Model (Carbaugh, Marzano, & Toth, 2017). The most recent update includes a standards-based approach to teacher evaluation, containing 23 teacher behaviors each falling under one of four domains. Marzano identified the four domains as the areas of expertise necessary for the teacher to be proficient to positively impact student achievement (Carbaugh, et al., 2017). The first domain, Standards-Based Planning, includes three elements related to the teacher's use of data to plan lessons and aligned the lessons to state standards. The second domain, Standards-Based Instruction, includes ten elements related to specific strategies used to determine content taught and the process used to teach. The third domain, Conditions for Learning, includes seven elements related to the assessment of student progress, student

feedback engagement, and classroom management. Finally, the fourth domain, Professional Responsibilities, includes three elements related to following district policies, collegiality, and staying current with educational research.

While classroom observation is essential to the Marzano Focused Teacher Evaluation Model (Marzano Model), two key activities that must also take place are the pre-planning conference and post-observation conference with the teacher. The pre-planning conference provides the teacher the opportunity to share specific strategies for the evaluator to observe. The post-observation conference gives the teacher the opportunity to provide student evidence, which demonstrate the impact the strategies had on student achievement, especially if the evaluator was unable to observe specific strategies while in the classroom. (Carbaugh, et al., 2017).

Scoring for the Marzano Model is competency-based and has five levels for each of the 23 elements in the four domains (Carbaugh, et al., 2017). It is not expected that all 23 elements be scored during each evaluation, but recommended the evaluator score all 23 a minimum of one time during the year. The five levels of scoring are: Not Using (0), Beginning (1), Developing (2), Applying (3), and Innovating (4). Because the Marzano Model is competency-based, and the goal is teacher growth leading to student achievement, the teacher's scores on the elements should increase with each observation on which that element is scored (Carbaugh, et al., 2017).

Applying these scores occurs during a specific five-step process outlined for the evaluator. This increases inter-rater reliability. In step one, the evaluator would look for the element or strategy the teacher indicated would be used during the pre-observation conference and determine if the strategy was used correctly. If the element is used

correctly, the evaluator would mark the teacher with a score of Developing (2) then move to step two. In step two, the evaluator would determine how the teacher is monitoring the students for learning as a result of the strategy used during the lesson. How the teacher monitor's the learning of the students does not change the score, however it does provide information for feedback to the teacher. In step three, the evaluator moves from teacher observation to evidence of student learning to determine the percent of students on whom the strategy is having the desired impact. It is necessary for the evaluator to observe student work to make this determination. The score applied at this point depends on the percent of students with the desired level of achievement. If that number were 50 percent or less, the score would remain a Developing (2). If the number were 51 percent to 90 percent of students, the evaluator would mark the teacher with a score of Applying (3). If the number were greater than 90 percent, the evaluator would mark the teacher with a score of Innovating (4). In a situation where the teacher does not score either a three or four, the evaluator would move to step four. In step four, the evaluator observes the teacher to determine if an adjustment is made to reteach in an effort to increase the number of students learning the skill. If the evaluator observes the adjustment increased the percent of students meeting the desired level of achievement, the teacher's score would be adjusted to reflect the evaluator's observation. If not, the score would remain the same. In step five, the evaluator would look at student evidence to assign a final score to the teacher. A final score may not always be necessary, as the evaluator may have been able to do this during the evaluation. If the evaluator was unable to give a final score, this step can be completed in the post-observation conference, with the teacher bringing evidence of the students' achievement (Carbaugh, et al., 2017).

As a competency-based model, the Marzano Model does not provide one final score by averaging all the scores together (Carbaugh, et al., 2017). The goal of the Marzano Model is the teacher will be provided opportunities over the course of the year, and in subsequent years, to improve through coaching, practice, and professional development. With this in mind, progress toward each of the 23 elements is tracked to determine the teacher's progress toward the mastery of each element. Low scores in the beginning, especially for new teachers, are expected. The scores offer evaluators the opportunity to provide the right feedback and encourage a growth mindset in the teachers. It is expected that future scores on those same elements will increase based on the feedback, coaching, and professional development received (Carbaugh, et al., 2017).

The Gates Foundation has conducted research regarding teacher evaluation and student achievement as the Measures of Effective Teaching (2010). The Measures of Effective Teaching provides links between teacher evaluation and student achievement, but it was self-published and funded. Additionally, the research has not been subjected to a review by other researchers for tests of validity and therefore was not used for the purpose of this research study (MET, 2012). To apply the theories that were chosen, a brief review of the history of teacher evaluation is provided in the next section.

History of Teacher Evaluation in the United States

Evaluation of teachers during the 1700s was generally the responsibility of either local government, or most likely clergy because of their extensive education; however, it was vastly different than present day evaluation (Marzano, Frontier, & Livingston, 2011). Marzano, et al. (2011) pointed out that community leaders, including clergy and business owners, typically set up schools and established criteria by which teachers would be

judged. In 2005, Burke and Krey (as cited in Marzano, et al., 2011) noted, because there were no common expectations of teachers across the country. During this time, evaluators could set individual criteria and had absolute control over every aspect of the evaluation process, including the employment of the teacher, and had the power to immediately dismiss a teacher if the evaluator did not feel the teacher met the specific criteria.

As the Industrial Revolution spread during the 1800s, the United States experienced population growth in urban locations, which brought about the development of school districts with common ideas about education, serving larger populations of students (Marzano, et al., 2011). The growth in public education made it apparent teachers needed to have the training in both the subject area being taught and the pedagogy of teaching (Marzano, et al., 2011). As the movement continued into the mid-1800s, teacher evaluation began to shift, with the focus on how the teacher could improve instruction; therefore, evaluators needed knowledge in the subjects being taught and the pedagogy of teaching as well (Marzano, et al., 2011; Tracy, 1995; Blumberg, 1985). Tracy (1995) noted during this movement was the emergence of a more formal system including a hierarchy that included school district superintendents, principals, and teacher trainers. Notable during this time, according to Blumberg (1995), was the realization on the part of schools that the quality of education was too elaborate a system to be evaluated by individuals not involved in education. Blumberg (1995) as well as Tracy (1995) went on to point out there was a common desire among public schools to provide a quality education and focus on training teachers accordingly.

As the 19th century ended and the 20th century began, John Dewey and Frederick Taylor offered two very different views on education (Marzano, et al., 2011). John

Dewey (as cited by Marzano, et al., 2011) maintained a view of education that mirrored our democratic society and advocated for democratic ideals to drive education. Dewey (as cited by Marzano, et al., 2011) believed students should be fully involved in their education. On the other hand, Frederick Taylor's view was scientific in nature (Marzano, et al., 2011). Taylor believed the best approach was to discover a way to perform a task, similarly done in a factory, and then choose the approach, which worked best (Marzano, et al., 2011). Taylor's approach viewed schools in much the same way as a factory, with the students seen as the product, and was favored by college professors who adopted the approach for use in specific courses (Marzano, et al., 2011).

In the early 1900s, psychologist, Edward Thorndike and educator, Ellwood Cubberley led educators to apply Taylor's scientific approach of measurement to the teacher evaluation process (Marzano, et al., 2011). Cubberley (1929) equated schools with factories and students with the product being manufactured. Additionally, Cubberley (1929) believed evidence of teaching should be obtained through data collected from observing the teacher. In the third edition of his book, *Public School Administration* (1929), Cubberley advocated for applying letter grades to the teacher based on performance. In 1929, William Wetzel added to Cubberley's observations by suggesting the evaluator use student data as an addition to observing a teacher's specific practices (as cited in Marzano, et al., 2011). Wetzel did not align his suggestions to Cubberly's model of comparing a school to a factory. Instead, Wetzel came up with three criteria for a scientific evaluation: student achievement on an aptitude test, objectives for each class which can be measured, and an accurate measure of student learning.

During the mid-1900s, there was a movement away from evaluation as a scientific process, and a shift toward helping teachers develop expertise as well as focusing on the emotions of the teacher (Marzano, et al., 2011). During this time, the responsibilities of the evaluator increased. While observation and evaluation of the teacher was on the list of evaluator responsibilities, they did not receive as much attention as many of the other items listed (Melchoir, 1950; Swearingen, 1946; Thompson, 1952, as cited in Marzano, et al., 2011). Despite the increased additional responsibilities of the evaluator, classroom observations and follow-up conversations were determined to be important in the evaluation process and added to the list of evaluator responsibilities (Whitehead, 1952, as cited in Marzano, et al., 2011). Tracy (2015) pointed out the apparent contradiction between the importance of culture and collaboration with teachers increased during this time, as did the need for stronger teacher evaluation systems. Tracy's may have resulted in supervisors' reluctance to be too harsh toward teachers (Tracy, 2015).

In the 1950s, Morris Cogan, a Harvard Professor, and his colleagues developed a system to use while supervising student teachers (Cogan, 1973, as cited in Marzano, et al., 2011). The resulting system was a five phase, clinical approach to the supervision of teachers that Cogan later wrote about in his book *Clinical Supervision* (1973). Cogan's goal was to provide a system the evaluator could work through with the teacher to assist the teacher in improving instruction in the classroom (1973). Cogan's process was eventually narrowed down to a three-phase process which included a conference with the teacher to discuss elements of the teacher's classroom and anticipated lesson, the observation of the lesson, and a conference following the lesson to discuss the evaluator's observation notes (Holland & Garman, 2001). A key part of evaluation was the

collaborative process of the teacher and evaluator working together to review the teacher's performance (Cogan, 1973, as cited in Marzano, et al., 2011). Cogan's evaluation model continued to be used until the early 2000s (Holland & Garman, 2001).

Charlotte Danielson's work with pre-service teachers led to a popular model of teacher evaluation, which became popular in the 1990s, and continues to be used by public school districts today. Danielson's Framework for Teaching included four domains: planning and preparation, classroom environment, instruction, and professional responsibility (2007). While the framework provided expected behaviors within the domains, it was not intended to be a checklist for evaluators, but rather a self-reflection tool for the teacher, through which the teacher could identify areas of strength and weakness and improve their craft (Benedict, Thomas, Kimerling, & Leko, 2013). Though Danielson's Framework for Teaching began as a model to assist in evaluating pre-service teachers, the model became a standard for many in education with its focus on the development of a common language for educators, a structure for evaluating and reflecting on professional practice, and recognition of the complex nature of the teaching profession (Danielson, 2007).

The publication of "A Nation at Risk" in 1983 was difficult for a nation that believed itself to be advanced in education when compared with its competitors. The publication was a catalyst to the increased focus on research into education and evaluation of teachers in the late 20th century. The United States began to take notice, and make decisions leading to educational reform (A Nation at Risk, 1983; Strong, et al., 2011).

Evaluation Reform

At the beginning of the 21st Century, government mandates and educational research resulted in the need for change in the evaluation of teachers (Strong, et al., 2011, Superfine, Gottlieb, & Smylie, 2012). The following is a review of government mandates for improved education and efforts to incentivize states to make changes to help improve student achievement. In addition, a review of research focusing on more effective evaluation processes to distinguish between good and poor teachers is included.

Robert Marzano (2012) identified two purposes for teacher evaluation. The first, measuring teachers, is a way to determine job retention. The second, developing teachers, is a way to provide opportunities for growth (2012). Marzano pointed out the significance of each purpose is very different; therefore, an evaluation for each purpose will differ (2012). Marzano conducted a survey of educators and found that 76 percent of more than 3,000 educators believed development of educators was the major goal of the evaluation process (2012). The goals of improving instruction and identifying ineffective teachers were also identified by Donaldson (2013) as purposes of teacher evaluation. In Donaldson's survey of principals, greater than two-thirds of the participants did not believe current evaluations were effective at improving instruction or identifying ineffective teachers.

Until the mid 1900s, the federal government had honored the role of states as the decision maker regarding education; however in 1954, the decision of the Supreme Court in Brown v. Board of Education brought about a new role for the federal government (Superfine, Gottlieb, & Smylie, 2012). After this landmark civil rights decision, declaring the segregation of students in schools based on race unconstitutional, Congress began to

pass laws and provide funding to protect the civil rights of citizens. In 1958, Congress passed the National Defense Education Act. By the 1990s, the government began to focus on teacher quality and student achievement, and passed the NCLB initiative in 2002 (Superfine, et al., 2012).

In 2002, NCLB created a sense of urgency for school districts across the country. The NCLB legislation required school districts to adopt standards and develop standardized assessments to measure those standards. Further, according to NCLB, school districts receiving Federal Title 1 funding were required to meet Adequate Yearly Progress (AYP), as outlined by NCLB. Failure to meet AYP resulted in a series of graduated sanctions or steps to follow to make improvement. The NCLB created a need to ensure teachers were effective in the classroom (Strong, et al., 2011).

In 2009, school districts were further challenged to increase student growth when the Obama administration introduced the Race to the Top initiative (RTTT). The Race to the Top initiative was designed to provide funding to school districts showing an increase in student achievement, with the purpose of serving as examples to school districts struggling to demonstrate growth. Considered a competition for funding, districts participating in RTTT earned points for meeting requirements. Various requirements included the creation of a performance based evaluation of teachers and principals based on multiple measures of effectiveness, adoption of a common set of standards to drive instruction, improvement of student performance on standardized assessments among students in schools performing at the lowest levels, protection of the development of charter schools, and development and use of data systems. Both the NCLB and the RTTT concerned school districts across the United States. Strong, Gargani, and Hacifazlioglu

(2011) pointed out, as a result of NCLB and RTTT, school districts had to begin to look at the effectiveness of the teachers and the ability to provide the environment and education students need to become proficient, as defined by the states. Darling-Hammond explained that policy makers were placing a great deal of emphasis on the need to improve student achievement, which led to discussions around teacher evaluations (2013). States hoping to receive money from NCLB Flexibility Waivers or RTTT took on the task of improving teacher evaluation systems and requiring their public school districts to implement higher evaluation standards (Darling-Hammond, 2013). States like Missouri felt the urgency to begin to develop a comprehensive evaluation system that examined not only a teacher's skills and abilities, but also at the growth of the student toward proficiency, as measured through annual standardized testing.

In 2015, President Obama reauthorized NCLB and renamed it the Every Succeeds Act (ESSA) with a few significant changes (Every Student Succeeds Act, 2017). Most notable in these changes was the move from a federal accountability system to a state accountability system. Under NCLB, the federal government placed requirements on the state education departments to carry out requirements regarding teacher evaluation, student assessment, and the adoption of standards. Under ESSA, the state was given more control in each of these areas. In regard to teacher evaluation, NCLB required states to use teacher evaluation if they were to receive waivers under NCLB. ESSA eliminated the requirement that states had to use federal funds for teacher evaluations. According to NCLB, assessments were mandated for all students in third through eighth grades and at least one time in high school. ESSA again gave states the choice in the administration of assessments, allowing states to administer the assessments in each specific grade and at

appropriate times. While NCLB wanted states to adopt college and career ready standards, ESSA mandated that the federal government remain neutral, allowing states to select standards (Every Student Succeeds Act, 2017). Because of the renewed ability to make decisions regarding education in each respective state, education departments began to look at current evaluation systems and assessments to determine how student achievement could be improved.

While states have been working to implement new evaluation tools to include value added measures, there remains little agreement on how to accomplish this goal (Grissom & Loeb, 2017). In their study of ratings from principals on teacher evaluations, Grissom and Loeb (2017) investigated the difference in evaluation scores given in both high-stakes and low-stakes situations for the same set of teachers. The high-stakes evaluation was a formal, summative evaluation required by the school district and used to make personnel decisions, which could result in termination. The low-stakes evaluation was an interview conducted in the spring, along with ratings from one to six, in eight job performance categories. The ratings from the high-stakes and low-stakes evaluation for each teacher were then compared. The findings showed principals gave scores of effective and very effective to greater than 97 percent of the teachers when evaluating teachers with the high-stakes evaluation. While the difference in low-stakes evaluation results were not significant, principals were more likely to give a lower rating in a low-stakes evaluation. In some instances, teachers receiving a score placing them in the ineffective category on the low-stakes evaluation received a score placing them in the effective or highly effective category on the high-stakes evaluation. Grissom and Loeb (2017) point out that, while value-added measures may be able to provide evidence of

some component of teacher performance, these measures are not able to provide evidence of all the components of teacher performance, which may impact student achievement.

Researchers have expressed concern with using student achievement, growth measures, and VAMs as components to evaluate teachers. Pressing among the concerns of the researchers is the inability to isolate the variables, which have led to student achievement and identify which are within the teacher's ability to control in the classroom. Factors including school attendance, socio-economic status, family involvement, and student motivation all play a role in student achievement, but they are outside the teacher's sphere of influence (Hattie, 2009; Ritter & Shuls, 2012; McDonnell, 2013; & Galey, 2015). Similar thoughts are echoed by Baker, Oluwole, & Green (2013) in their study regarding the use of VAMs in teacher evaluation. Baker, et al. (2013), go a step further, indicating the unreasonableness, and possible violation of a teacher's due process rights, when trust is placed in an unproven method to evaluate teachers, especially when many other factors may influence the outcome. Another concern expressed by educational researchers regarding the use of student achievement, growth measures, and VAMs to evaluate teachers is the ease with which the data can be misused, which can lead to inaccurate conclusions (Senechal, 2013; Koyama & Kania, 2014). Papay (2012) does offer a solution, suggesting the use of multiple years of data for evaluative purposes as opposed to a single year's data.

In most states, standardized tests are used for the purpose of teacher evaluation scores in regard to student achievement whether related to growth or a value-added measure. The issue standardized tests present is a limited view of teacher effectiveness and accountability (Brevetti, 2014). As pointed out by Turnipseed and Darling-Hammond

(2015), when such a limited view is presented, it is not possible to provide teachers with the support necessary to make improvements that will impact instruction or student achievement, thus creating a system of little to no worth, ineffective, and useless (Piro & Mullen, 2013). Feingold (2013) found in employing the ineffective practice of using standardized test scores to evaluate teachers, states stand to harm rather than help the education profession. Feingold (2013), along with Jennings & Sohn (2014) suggest a more collaborative approach to evaluation. The suggested approach would work between the teacher and evaluator and consider the many facets of the teacher's responsibilities, by taking into account the day-to-day functions and relationships required (Mausethagen, 2013). Additionally, Jacob (2012) suggested the inclusion of elements that encourage frequent reflective practice by the teacher would lead to improved instruction.

Educators, including teachers, agree there is a need for effective evaluation systems (Marzano, 2012). However, despite the continued changes and reforms to the evaluation systems, criticisms continue. Marzano identified one limitation as the inability of the evaluation system to distinguish effective teachers from ineffective teachers (2012). When the evaluation does not identify whether the teacher is effective, it cannot give an accurate measure of the quality of instruction or help to improve instruction (Measures of Effective Teaching, 2010; Marzano, 2012). Sandilos, Sims, Norwalk, and Reddy (2019) examined the data from three evaluations used by the Gates Foundation to gain insight on what measures should be used to evaluate teachers. The evaluations included the Classroom Assessment Scoring System (CLASS), Framework for Teaching (FFT), and the Tripod. The CLASS proved to be most effective at the elementary school level, while the Tripod, a student evaluation of the teacher, proved to be most effective at

the middle school level. When predicting VAMs, high scores in classroom management proved most accurate in the area of math, while warmth and rigor were the areas where high scores proved most accurate in English language arts (Sandilos, et al., 2019).

In a study regarding evaluation practices in Chicago Public Schools, questions were raised regarding the impact of the climate and environment of the school on the teacher's ability to impact student achievement (Sporte, Jiang, Luppescu, & Society for Research on Educational Effectiveness, 2016). Teachers in schools with a large population of high poverty students received evaluation scores significantly lower than schools with fewer students from poverty, even when value-added scores were not significantly different. Teachers working in schools with a solid professional climate tend to receive higher evaluation scores. Teachers from minority groups, as well as male teachers, received lower evaluation scores, even with comparable experience. As it relates to minority teachers, value-added scores were not significantly different. While the value-added measures may be able to identify teachers who are using effective teaching strategies, these results lead to questions about whether the evaluation used is implemented effectively, based on what the teacher is able to do in the classroom, and not on outside factors. Good teachers, working in schools with high poverty or a poor professional climate may not receive a fair evaluation. Additionally, teachers may seek employment where it is easier to get a fair evaluation (Sporte, et al., 2016).

Classroom observations are an essential component of teacher evaluation models (Whitehurst, Chingos, & Lindquist, 2015). Whitehurst, et al. (2015) conducted a study of four schools making progress in effective teacher evaluations by placing emphasis on the classroom observation experience. The researchers made three recommendations

following the study. First, the researchers recommend teachers be observed two or three times each year, with one of the observations completed by an evaluator from outside the school. The completion of an observation by an outside evaluator ensures unbiased review of the teacher. Second, the researchers recommend the observation have the same or higher weight of any VAM or growth, which may be applied. Finally, the researchers recommend observation scores be adjusted for teachers, based on the characteristics of the class. An example could include adjusting for a classroom with students from a lower socioeconomic background, as well as adjusting for a class whose students are higher achieving and will therefore not show as much growth (Whitehurst, et al., 2015).

In the 2009 study, *The Widget Effect,* Weisberg, Sexton, Mulhern, and Keeling challenged teacher evaluation methods, positing that it was rarely used as a growth tool and most often used as a means to dismiss a teacher. In the study, Weisberg et al. (2009) gathered information from close to 15,000 teachers and 1,300 administrators in 12 districts in four states. The findings of Weisberg et al. (2009) included a consistent rating of good or satisfactory for the majority of teachers, with 94-99% of the teachers receiving the highest rating, a lack of recognition for teachers who were exemplary due to the consistent high ratings given to those who were even less than satisfactory, a lack of attention to newer teachers who also received good or satisfactory ratings despite their lack of experience, and a lack of attention to the performance of teachers who were ineffective, but continue to receive satisfactory ratings. The study went on to note that the situation was made worse because the process for evaluations was typically based on no more than two observations of no more than 60 minutes in total (Weisberg, et al., 2009). Weiseberg, et al. (2009) also indicated the evaluations were often conducted by

administrators who were not trained, as well as in a culture where the teacher expected to be highly rated based on past evaluation practices. Weisberg, et al. (2009) believed implementing the following recommendations could reverse *The Widget Effect*:

1. Adopt a comprehensive performance evaluation system that fairly, accurately, and credibly differentiates teachers based on their effectiveness in promoting student achievement.
2. Train administrators and other evaluators in the teacher performance evaluation system and hold them accountable for using it effectively.
3. Integrate the performance evaluation system with critical human capital policies and functions such as teacher assignment, professional development, compensation, retention, and dismissal.
4. Adopt dismissal policies that provide lower-stakes options for ineffective teachers to exit the district and a system of due process that is fair but efficient. (pp. 27-40).

Nearly all school districts across the United States acted on the recommendations of Weisberg (2009) and his colleagues, implemented the criteria outlined to reverse *The Widget Effect*. In 2017, Kraft and Gilmour reported in a study published in *Educational Researcher* that little had actually changed in teacher ratings. Kraft and Gilmour (2017) further noted in their findings, while some small change could be seen in the categories above and below the proficient rating, the percent of teachers receiving the rating of unsatisfactory had remained unchanged.

Some in the field of education view the teacher evaluation process as simply a way to meet the requirements of bureaucracy, and not a true tool for the development of

teachers (Maslow & Kelley, 2012). The creation of NCLB and RTTT, placed states in the position of rushing to respond, causing principals uneasiness about the lack of preparation needed to effectively implement the new evaluation (Sadeghi & Callahan, 2013; Kirkpatrick, 2010). Sawchuk (2013) believed the lack of adequate training forced evaluators to assign inflated scores to teachers.

Considering the reform mandates required under NCLB, and research regarding best practices related to evaluation, changes were needed. The following is a review of three of the evaluation models approved by MODESE for use by school districts in Missouri. Each of these models follows MODESE's principles for an effective evaluation.

Missouri's Teacher Evaluation

Missouri's response for improved student achievement included requiring school districts to adopt an approved evaluation system (MODESE, 2015). Districts could choose from the Model Educator Evaluation System, which was developed by DESE, the Network for Educational Excellence, a system developed by Missouri University, or could choose to create their own model that followed the principles identified by DESE. Each of the models are described in the following section (MODESE, 2013c).

Information related to the MAP assessment and the Missouri MEES was obtained for Chapter Two from the MODESE website. The researcher searched for MODESE source material to locate the peer-reviewed sources MODESE used to develop the MAP and the MEES. An extensive search of MODESE failed to produce a clear explanation of the research used to develop the MEES, only a list of research, which had been considered in its development. In an effort to determine if other researchers had

researched the development of either the MAP or the MEES, the researcher searched academic databases, but these searches did not yield results. However, the researcher chose to move forward, using MODESE as the source for the MAP and the MEES for the purpose of Chapter Two.

Model Educator Evaluation System. As a result of a flexibility waiver received for exemption from NCLB, Missouri DESE piloted a new evaluation tool during the 2012-2013 school year known as the Missouri Model Educator Evaluation System (MEES) in more than 100 school districts in Missouri (MODESE, 2013c). Like many other states, Missouri responded to the need to effectively evaluate educators in an effort to provide the best education for students, with the hope of ensuring Missouri school districts are meeting the guidelines set forth by NCLB. The goal of the evaluation was to improve teacher quality with the outcome of increasing student achievement (MODESE, 2015). All Missouri school districts are required to implement the Model Educator Evaluation System evaluation tool, the Network for Educator Excellence, or a tool created by the district meeting the principles set forth by DESE, on which the Missouri Model Educator Evaluation System is based (MODESE, 2014).

Missouri's Educator Evaluation System is based on seven principles, which are outlined in DESE's Essential Principles of Effective Evaluation document provided on DESE's website (2013). The first principle is "Research-Based and Proven Performance Targets" (p. 2). The first principle outlines for the teacher and the evaluator a structure of clear expectations, which should be recognized in the classroom. The second principle is "Differentiated Levels of Performance" (p. 3). The second principle is also a structural component of the evaluation and identifies the areas for growth, based on the specific

needs of the teacher. The third principle is another structural component, "Probationary Period for New Educators" (p. 3). Teachers in their first five years of teaching and are placed in a probationary period, providing an adequate length of time to gain the skills necessary to be effective in the classroom. The fourth principle, related to the process of evaluation, is "Use of Measures of Student Growth in Learning" (p. 3). A common goal of most educators is for students to experience academic growth during the school year. The fourth principle ties student growth to the teacher and holds the teacher accountable for growth. Principle five, "Ongoing, Deliberate, Meaningful and Timely Feedback" is also related to the process of the evaluation (p. 4). The fifth principle requires the collaboration between the evaluator and the teacher be ongoing and based on the goals set for the teacher. Principle six, also related to the process of the evaluation, is "Standardized and Periodic Training for Evaluators" (p. 4). Evaluators are required to receive training to maintain reliable and valid measures of the teachers they evaluate. The final principle is "Evaluation Results to Inform Personnel Employment Determinations, Decisions and Policy" (p. 5). The seventh principle gives districts the opportunity to identify the highly effective teachers on staff and use those teachers to improve instruction school-wide, through mentoring, coaching, leadership, and other essential positions impacting the academic success of students.

The research used to develop the MEES is based on Robert Marzano's research on teacher evaluation. Marzano's strategies for the classroom were included in the teacher evaluation growth guide and aligned to the standards and indicators in an effort to support teachers' professional practice (Marzano, 2007, as cited in MODESE, 2012b). Because Marzano's strategies are research-based, teachers are provided with a toolbox of

strategies, which support, guide, and improve professional practice with the overall goal of improving student achievement (Marzano, 2007, as cited in MODESE, 2012b).

The Network for Educator Effectiveness (NEE). The University of Missouri's College of Education, in an effort to assist Missouri school districts in meeting Missouri DESE's requirements for teacher evaluation, invited experts on both assessment and professional development to design an evaluation system (NEE, 2017). The resulting evaluation, the NEE, is an online tool that provides supervisors with training to implement the NEE system, as well as a library of professional development videos teachers can view, based on feedback received from the evaluation (NEE, 2017). Though Missouri DESE created its own model, the NEE is used by more school districts in Missouri than any other evaluation system (2017).

District Created Evaluation Systems. While MODESE developed its own evaluation system based on seven principles Missouri identified as necessary for effective evaluation of a teacher, it does not require that the school district use the model. According to the Missouri Secretary of State's Code of Regulations, any Missouri school district may choose to develop an evaluation system. If the district opts to create their own evaluation system, the district must follow the seven principles outlined by MODESE as necessary to effectively evaluate a teacher. Additionally, the evaluation tool must be submitted to MODESE for review and approval prior to use (Mo. Rev. Stat. §168.128, 2014).

While MODESE has approved evaluation systems for Missouri school districts to implement, the impact on teacher growth leading to student achievement is unknown. Each of the evaluations, the MEES, NEE, and a district created evaluation must meet the

criteria MODESE has established and is required by NCLB and ESSA. As a result, evaluators will include a focus on Marzano's theory, which should enhance a teacher's pedagogy and improve student performance (Marzano, 2012). If the purpose of the evaluation is improvement of the teacher's skills and abilities for the purpose of increasing student achievement, the student evidence components of these evaluations will need to be effective in assessing teacher effectiveness (Marzano, et al., 2011). Therefore, consideration should be given to factors which impact student achievement.

Student Achievement

As a result of NCLB, states have included student growth, which is tracked from state assessment, as a component of the evaluation tools to measure teacher effectiveness (Marzano, 2013). While there is agreement among researchers that the inclusion of student growth is important in the evaluation of teachers, there is some disagreement about how the growth should be measured. Regardless of the measure used by the evaluator, consistency and an understanding of the influences on student achievement are necessary for the evaluator to make a fair judgment when conducting an evaluation (Marzano, 2013). The following is a review of student achievement and the factors which influence student achievement.

Influences On Student Achievement. John Hattie and Eric Anderman (2013) asked experts in the field of student achievement to contribute to their *International Guide to Student Achievement.* Each author shared his or her thoughts and expertise regarding student achievement. In his chapter of the book, Thomas Guskey defined achievement as "the accomplishment of something" which is typically based on learning that occurs in the school environment (Hattie & Anderman, 2013, p. 3). Guskey pointed

out that schools can agree student achievement is the basis for decisions in education; however, there is little to no agreement on what student achievement actually is or looks like (Hattie & Anderman, 2013). Guskey went on to explain that achievement can be viewed and measured in many different ways and continues to change and evolve as a student advances in school. Subject areas measure student achievement through various student work products, which can include anything from exams and tests to written research projects to oral presentations. Achievement can be seen as acquisition of new information or improving knowledge or knowledge already obtained. Achievement is influenced by many factors, both known and unknown. (Hattie & Anderman, 2013).

Guskey added the definition of student achievement is further complicated by two other factors (Hattie & Anderman, 2013). The first factor is determining whether student achievement is the attaining of new information, or the improving of information previously attained. Attainment refers to a student's ability to achieve at a particular level and how that student compares with other students the same age, in the same grade level, during the same time. Students in this category are typically referred to as proficient or on grade level. Improvement refers to a student's growth, or change, in previous learning. Improvement is based on pre- and post-test results to determine how much has been learned. While achievement and improvement are connected, a student may make significant improvement; yet never meet a level of proficiency or grade level. Additionally, a student may show significant achievement and a high level of proficiency above grade level in comparison with peers; yet show very little growth (Hattie & Anderman, 2013).

The second factor Guskey referred to, which complicates the definition of student achievement, is instructional sensitivity (Hattie & Anderman, 2013). An assessment that is instructionally sensitive seeks to assess a student's proficiency related to the objectives being taught in a particular unit of study. If an assessment is instructionally sensitive, it is effective at assessing only the knowledge the student should have obtained following what was taught. If the majority of students score well on the assessment, the teacher can surmise he or she did well teaching the subject (Hattie & Anderman, 2013). Instructional sensitivity would support the theory of rationalism, as it is the logical if-then thought process. In addition, the teacher who believes he or she did well teaching a subject based on the results received from a previous administration of an assessment, is likely to repeat the same pattern of behaviors expecting the same results, and is unlikely to alter even if the results change (Bhola, 1990; Darling-Hammond & Wise, 1981).

Hattie and Anderman (2013) asked Alan Bates, Rena Shifflet, and Miranda Lin to write about student achievement as it relates to elementary school students. According to Bates, Shifflet, and Lin, achievement for elementary school students is typically measured through reported grades on grade cards or standardized assessment results (2013). Reported grades on grade cards and standardized assessment results are problematic. Traditional letter grade systems vary greatly from teacher to teacher, and are subject to each teacher's expectations of the students and standards set for the teacher's class; therefore, the same grade in the same subject and grade level class may not indicate the same level of achievement or knowledge because of the subjectivity of the individual teacher's grading practices. Standardized tests, on the other hand, are created to eliminate

subjectivity, yet still pose issues in determining student achievement because of the inability to assess higher-order thinking and understanding (Bates, Shifflet & Lin, 2013).

Many factors can impact an elementary student's achievement. The student's intelligence has a large impact on achievement, as does motivation. Student achievement can also be predicted based on the influence of "family, socioeconomic status, community, and school" (Bates, et al., 2013, p. 7). While a student's intelligence and level of motivation are both great predictors of achievement, studies have indicated the student's social environment, including family, community, and home can have a significant impact on the student's achievement. If the social environment is not positive and supportive, the student may not reach the level of achievement indicated by the level of intelligence he or she exhibits (Emory, Caughy, Harris, & Franzini, 2008).

Parental involvement in education and school attendance are two more factors important in the achievement of elementary students. When parents are involved in their child's school through volunteer activities as well as encouraging study habits at home, student achievement increases (LaRocque, Kleiman, & Darling, 2011). Consistent attendance in school is also an important factor in student achievement. Research has shown students with higher achievement in elementary school tend to have higher attendance rates when compared with students with lower achievement (Bates, et. al., 2013).

Teachers are also a large factor in student achievement. According to Hattie (2009), teachers and teaching can have the largest impact on student achievement. In Hattie's meta-analysis of factors impacting student achievement, four of the top twenty were attributed to the teacher and eight were attributed to the teaching, all having an

impact of over one year of growth in student achievement (2009). Pil and Leana (2009, as cited in Bates, et. al., 2013) found teachers who collaborated with their teaching partners, as well as those who had experience in a particular grade or discipline tended to have students who obtained higher academic achievement in comparison to students who did not have teachers with these backgrounds. In her 2017 study of the relationship between teacher performance and student achievement, Parker determined there is a relationship between teacher performance and student achievement in language arts, and a weak relationship between teacher performance and student achievement in math. Parker (2017) did note that the teachers who focused on a specific subject area tended to get higher results, particularly in the area of math.

Measuring student achievement. The Missouri Assessment Program (MAP) was Missouri's answer to the requirements of the 1993 Outstanding Schools Act (MODESE, 2014). The Outstanding Schools Act required Missouri to develop standards students should learn before graduation, as well as an assessment to measure acquisition of the standards. The MAP was designed with two purposes; to determine how well students attained the skills outlined in the Missouri Learning Standards (MLS) to meet Missouri's goal of graduating students who were college and career ready, as well as to determine the quality of the education students in Missouri were receiving (MODESE, 2014). When NCLB was passed in 2001, the MAP was used to report student proficiency levels to ensure Missouri, its schools, and districts were making adequate yearly progress (AYP). Throughout the years, the MAP has gone through a few changes and is currently an online assessment given to students in grades three through eight in English language arts and math, and science in grades five and eight (MODESE, 2014).

Missouri brought together a group of citizens, including parents, educators, and business people, to develop the Show-Me Standards, on which Missouri education is based (MODESE, 2014). The 73 Show-Me Standards, 40 knowledge standards and 33 process standards, were developed with the intention of providing districts a foundation on which to base the development of a curriculum. While the standards are not all inclusive, districts were able to use them to develop a challenging curriculum that prepared students for life post-graduation (MODESE, 2014). To aid districts in developing curriculum, the Missouri Learning Standards (MLS) were developed with assistance from Missouri educators (MODESE, 2014). The MLS outline the minimum, specific learning requirements, which should take place at each grade level, in order to ensure students move toward acquisition of the Show-Me Standards before graduation. The standards are not a curriculum, however do assist educators in determining what to teach and are the basis for items on the MAP (MODESE, 2014).

Missouri requires the administration of the MAP annually in the spring (MODESE, 2014). In grades three through five, the MAP is administered in English language arts and math, and science in grade five. While students are not permitted to take a test outside the current grade level, Missouri has created an alternate assessment for students with severe cognitive disabilities. The MAP-Alternative is given in English language arts and math in grades three through five, and science in grade five for only those students who have an Individualized Education Plan (IEP), and qualify due to a severe cognitive disability (MODESE, 2014). Because of the limitation of the accommodations in MAP testing, the assessment can be considered a strong example of the Rationalistic Theory of student achievement, due to a simplistic view of student

achievement (Darling-Hammond & Wise, 1981). According to the Rationalistic Theory, there is a belief that a uniform set of goals will bring about the same results in all classrooms, a belief all teachers will deliver instruction in the same way, and a belief all students will learn in the same way (Darling-Hammond & Wise, 1981). The MAP is one test, administered in basically the same way to all students unless students have an Individualized Education Plan (IEP). All teachers are required to administer the same test across the state and all students take the same test, regardless of their learning style or preference. According to Darling-Hammond and Wise (1981) the Rationalistic Theory also pointed out that the student is not seen as an active participant in their learning, which the MAP supports by administration of the assessment in late spring and delivery of the results sometimes as late as the fall of the following school year.

A variety of items are presented on the assessment, each with benefits and drawbacks in the ability to assess the students' achievement. Multiple-choice items give the student a question, followed by several choices from which the student must choose the correct answer. These items demonstrate what each student knows and understands and are easy to administer and grade. Because of this, test administrators can ask many questions and get a measure a broad spectrum of information students may have learned. A limitation of multiple-choice items is the inability to measure a student's ability to apply the knowledge learned. Technology enhanced items are computer-based items, which require a response by computer. In addition to allowing the student to demonstrate what has been learned, technology enhanced items "allow students to demonstrate what they know in an authentic way" and the students benefit from computerized scoring (MODESE, 2014, p. 6). Constructed response questions require the student to provide an

answer to a question. The benefit of this type of question is it can require the student to show how he or she arrived at the answer. Constructed response questions are limited in their usage due to the time it takes to both administer and score them. The final type of question on the MAP is the performance event, which requires the student to apply what has been learned. Performance event questions may require the student to spend a large amount of time, complete multiple steps, and draw from several standards learned to answer the question. While this type of question is excellent at measuring a student's knowledge and understanding of standards as well as the ability to apply the learning to a real-world situation, it is incredibly time consuming and costly to both administer and score (MODESE, 2014).

The reauthorization of the Individuals with Disabilities Education Act (IDEA) in 2004 required the inclusion of special needs students in state accountability measures. The inclusion of special needs students included students receiving special education services, students with disabilities, and English language learners (MODESE, 2014). In 2014, Missouri provided for all populations of students when the MAP moved to an online assessment. Universal tools were provided for all students and included aids like a highlighter, spell check, and a ruler. Designated supports were provided for only those students for whom an educational team determined the tools were needed and included aids like color overlays, magnification, and a scribe. Accommodations were supports provided only to students who had an IEP or a 504 plan that specifically identified one of the supports as a need for the student. These supports included American Sign Language, multiplication tables, and adapted keyboards (MODESE, 2014). Overall, the changes that

affected student achievement also impacted teachers, their work, and their perception of these new expectations.

Teacher Perceptions of Evaluation Reforms

Teacher perception studies provide insight into how teachers have responded to the new expectations as a result of the new evaluation models, specifically the addition of growth measures and VAMs. Graziano (2017) studied teacher perceptions of Marzano's model. Her findings showed teachers overwhelmingly believed they used the strategies Marzano outlined in his model. The teachers further reported they may change their practice to comply with the expectations of the model but not to affect student achievement. Teachers did appreciate the feedback from the model, and felt it provided them with information that could help them improve in some aspects of their instruction, but did not accept the model as an evaluation which could apply in all educational settings (Graziano, 2017).

In her 2019 study of elementary school teacher perceptions, Amy Long referenced the lack of research regarding new evaluations created following federal reforms and their impact on student achievement and teacher practice. Through surveys and interviews, Long discovered 66 percent of the teachers believed the new evaluation led to changes in their instructional practice. She further explored the specific feedback, which led teachers to make changes. Long identified five evaluation characteristics had an impact on how the teacher responded to the results of the evaluation. Long found teachers were more likely to make changes to their practice when the following conditions were present in their evaluation: the teacher was involved in setting the goal and believed the goal would result in student achievement gains, feedback from the evaluator was useful,

and there was a relationship of trust between the evaluator and the teacher. Teachers were less likely to make changes to their practice when the following conditions were present in their evaluation: evaluations were rushed or feedback was limited due to lack of time on the part of the evaluator and the perception of the teacher that the evaluation, despite the efforts of the district, is subjective and therefore not based on the teacher's skills or competency (Long, 2019). Additionally, an interesting point made by Galey (2015), despite the impact of evaluations on teachers and the high stakes placed on the teachers for personnel decisions, teachers are rarely included in the development or decisions as to what to include in the evaluation process.

In his 2018 study of teacher and principal perceptions of teacher evaluations, which considered perceptions of both the quality and accuracy of evaluations, Lewis found principals to have more confidence in the quality of the teacher evaluations conducted. Lewis found this difference to be statistically significant. However, while he found principals to rate the accuracy of evaluations higher than teachers, this difference was not significant. Lewis found four of the six areas used for evaluation to have a statistically significant difference despite the fact that each of the teachers included in the survey were rated as effective or highly effective. The four areas included evaluation for the purpose of professional development, retention of teachers, awarding tenure, and decisions regarding termination. The two areas principals rated higher than teachers, yet not statistically significant, were teacher compensation and advancement decisions (Lewis, 2018). One conclusion Lewis drew from his research was an apparent disconnect between what was observed and evaluated, what actually happens in the classroom every

day, and the inability of an evaluation tool to capture what happens accurately (Lewis, 2018).

Summary

Teacher evaluation has gone through many changes since the early 1700s. Extensive research into best practices as well as government reform efforts to increase student achievement have led states like Missouri to put effort into improving evaluation measures to include observations of teachers, feedback on improvement measures, and student growth and achievement (Weisberg, et al., 2009: Kraft & Gilmour, 2017; Maslow & Kelley, 2012). Standardized testing, such as the MAP given in Missouri, track student achievement over time to determine teacher effectiveness in teaching the standards adopted by the states. Missouri developed specific principles by which teachers are expected to be evaluated and requires each district in the state to choose from a limited number of evaluation models when implementing evaluation protocols, all in an effort to improve student achievement (MODESE, 2013a).

While the research indicates the teacher has the most influence on the achievement of students, teacher influence is difficult to evaluate. Traditional measures of student achievement, such as letter grades, tend to be subjective. Standardized tests attempt to take out the issue of subjectivity, however such standardized tests are unable to take into consideration the many other factors, which impact student achievement. Additionally, there are many different ways to view achievement, many ways to measure achievement, and many factors other than the teacher, which can influence achievement. Standardized tests are one view, one measure, and cannot consider the many influences on achievement, which may have impacted the student's score.

Chapter Three will present the research design and methodology for the study. The selection of the participants for the study will be presented as well as the research setting and design. Additionally, the procedures, instrumentation, and data analysis will be described.

Chapter Three

Research Design and Methodology

Introduction

This causal-comparative study was conducted to determine the difference in the percent of students scoring proficient and advanced on the third, fourth, and fifth grade Missouri Assessment Program, depending on the evaluation tool used to evaluate the teacher for the years 2017, 2018, and 2019 in the state of Missouri. Missouri school districts chosen for the study used the MEES, NEE, or a district created evaluation system approved by DESE for the years 2017, 2018, and 2019. A stratified, purposive sampling was used to select the school districts to include in the study. Data related to the percent of students scoring proficient and advanced for each of the school districts was retrieved from the Missouri Comprehensive Data System portal at the Missouri Department of Elementary and Secondary Education open access website. To determine if there was a difference in the scores, the researcher used a one-way ANOVA.

In this chapter, the research addresses the purpose of the study, including the research questions and the null hypotheses. The participants in the study are introduced followed by the sampling process. The research setting and research design are outlined as well as the procedures used. Finally, the instrumentation utilized and the data analyses are detailed as well.

Purpose of the Study

The purpose of this causal-comparative study was to compare the differences in the percent of students scoring proficient and advanced on the MAP in third, fourth, and fifth grades, depending on which evaluation system was used to evaluate teachers

between the years 2017 and 2019. The independent variable of the study is the evaluation system used by the school districts. The choices between the evaluation system used for the study include the MEES, the NEE, or a district-created evaluation system. The dependent variable of the study is the sum of the proficient and advanced scores on the MAP scores for the years 2017 through 2019.

The purpose of teacher evaluation systems is two-fold; supervision and evaluation (Mette, et al., 2017; Range, 2013). Supervision is typically seen as a support function, providing feedback to help the teacher improve, while evaluation is typically seen as the formal practice of holding the teacher accountable to state standards (Mette, et al., 2017). An effective teacher is essential to student achievement and will therefore have a strong impact on student achievement. In 2009, John Hattie provided a comprehensive look at research studies conducted to determine what had the largest impact on a student's academic success. Of the top 20 influences on student achievement, four were attributed to the teacher and eight were attributed to the teaching, all having an effect size of between 0.61 and 0.90. Teachers who have a positive impact on student achievement can actually lead to greater than a year's academic growth for a student. As a result of the teacher having such potential to impact student achievement, evaluators need effective tools to provide guidance and feedback which will help the teacher improve. What is not clear is whether the evaluation systems DESE has approved lead to greater student achievement. Despite many changes in evaluation systems across the United States in response to the Race to the Top and the NCLB's promised incentives, there was minimal change in the number of teachers determined to be ineffective (Dee & Wyckoff, 2017). Although Hattie (2009) likely provided the most comprehensive list of variables, which

improve learning, not all of the variables can be measured in a teacher evaluation.

In 2012, Missouri received a waiver from the requirements of NCLB, allowing the state the opportunity to identify its own system of accountability within the state's school districts, and focused on school improvement and student achievement (MODESE, 2015). One component Missouri promised as part of the waiver was the development of a teacher evaluation system, which would be more comprehensive, holding teachers accountable for student achievement (MODESE, 2015). This study examined the evaluation systems created to fulfill the requirements of the waiver, and used by school districts in the state of Missouri, to determine if there is a difference in student achievement in grades three, four, and five on the Missouri Assessment Program based on the evaluation used.

Research Questions

The following research questions have been developed to address the stated problem:

1. What are the differences in the **third grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system and DESE approved, district developed evaluation model in **2017**?
2. What are the differences in the **fourth grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Missouri Educator Evaluation System, the Network for Educator Effectiveness

evaluation system and DESE approved, district developed evaluation model in **2017**?

3. What are the differences in the **fifth grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system and DESE approved, district developed evaluation model in **2017**?
4. What are the differences in the **third grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system and DESE approved, district developed evaluation model in **2018**?
5. What are the differences in the **fourth grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system and DESE approved, district developed evaluation model in **2018**?
6. What are the differences in the **fifth grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system and DESE approved, district developed evaluation model in **2018**?

7. What are the differences in the **third grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system and DESE approved, district developed evaluation model in **2019**?
8. What are the differences in the **fourth grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system and DESE approved, district developed evaluation model in **2019**?
9. What are the differences in the **fifth grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system and DESE approved, district developed evaluation model in **2019**?

Null Hypotheses

The following null hypotheses were investigated to answer the research questions:

$H_0$1. There are no statistically significant differences in the **third grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system and a DESE approved, district developed evaluation model in **2017**.

$H_0$2. There are no statistically significant differences in the **fourth grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system and a DESE approved, district developed evaluation model in **2017**.

$H_0$3. There are no statistically significant differences in the **fifth grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system and a DESE approved, district developed evaluation model in **2017**.

$H_0$4. There are no statistically significant differences in the **third grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system and a DESE approved, district developed evaluation model in **2018**.

$H_0$5. There are no statistically significant differences in the **fourth grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system and a DESE approved, district developed evaluation model in **2018**.

$H_0$6. There are no statistically significant differences in the **fifth grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator

Effectiveness evaluation system and a DESE approved, district developed evaluation model in **2018**.

$H_0$7. There are no statistically significant differences in the **third grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system and a DESE approved, district developed evaluation model in **2019**.

$H_0$8. There are no statistically significant differences in the **fourth grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system and a DESE approved, district developed evaluation model in **2019**.

$H_0$9. There are no statistically significant differences in the **fifth grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system and a DESE approved, district developed evaluation model in **2019**.

Participants

Participants chosen for the study were school districts in the state of Missouri using the MEES, NEE, or a district created evaluation tool approved by DESE for use in the school district to evaluate teachers. Each district used the same evaluation system for all three years of the study. Each district served elementary students in grades three through five. Schools selected were from rural, urban, and suburban communities. A total

of 371 districts participated in the study and administered either the MEES, NEE, or their own district created evaluation system, and used the evaluation consecutively for each of the years 2017, 2018, and 2019. Of these schools, 46 used a district created evaluation, 86 used the MEES, and 241 used the NEE.

Selection/Sampling

For the purpose of this study, public school districts in the state of Missouri, with elementary students in grades three, four, and five were chosen. Stratified, purposive sampling was the method chosen to select the school districts to participate in the study. Stratified sampling is used when a researcher wants to ensure the appropriate representatives desired from a subgroup are included in the study (Gay, Mills, & Airasian, 2009). The stratified sampling is purposive because the schools selected for the study met the criteria set by the researcher (Gay, et al., 2009). To determine the sample, the researcher began by identifying which evaluation method each school district in Missouri administered for the years 2017, 2018, and 2019. The evaluation system implemented by each school district is located on the MODESE open access website in the Missouri Comprehensive Data System (MCDS) web applications. The list was then narrowed to only those districts that had used the same system consecutively for three years. The researcher included all districts meeting the criteria in the study. To make the comparison, it was determined, an alpha level equal to .05, a medium effect size, and a power of .08 would be sought (Gay, et al., 2009). Alpha represents the point at which it is determined an outcome cannot be attributed to chance, and is typically set at 0.05, or five percent. This is based on the work of statistician R. A. Fisher, who determined that an occurrence of once in 20 trials was rare enough to attribute statistical significance to an

outcome and reject the concern that the outcome could be due to chance (Bruce, 2015). Additionally, according to Gay, et al. (2009), effect size represents the strength of the relationship between the variables in the study. Effect size can have a negative or positive value and is expressed as a decimal. An effect size around .20 would indicate a small effect while an effect size around .80 would indicate a large effect (Gay, et al., 2009). For the purpose of this study, a medium effect size was sought. Finally, a power of .8 was chosen for the study. "Power refers to the ability of a significance test to identify a true research finding (i.e., there's really a difference, and the statistical test shows a significant difference), allowing the experimenter to reject a null hypothesis that is false" (Gay, et al., 2009).

Missouri Assessment Program data for each of the districts third, fourth, and fifth grades, for the years 2017, 2018, and 2019 was collected from the MCDS portal at the MODESE open access website. Ex post facto data was collected for the purpose of this study. Ex post facto data was appropriate for this causal-comparative study as ex post facto data is data previously compiled in some manner, however not necessarily for research purposes (Simon & Goes, 2013). Using ex post facto data provided the opportunity to use three years of data for the purpose of comparison to determine if any difference existed in MAP scores, when compared to the evaluation system used to evaluate teachers.

Research Setting

The school districts participating in this study had students in grades three through five and were located in the state of Missouri. The researcher chose a stratified, purposive sample of public-school districts in Missouri utilizing one of the following evaluation

systems for the years 2017, 2018, and 2019; the MEES, NEE, or a district developed evaluation system. The school districts chosen for the study varied in size and enrollment. Additionally, school districts varied in free and reduced lunch status, assessed valuation, and location in the state of Missouri. The districts were located in urban, suburban, and rural communities.

Research Design

The researcher chose a quantitative, causal-comparative design for this study. For the purpose of the study, the independent variables were the evaluation system used by each school. The independent variables included the MEES, NEE, or a district created evaluation system, which met DESE's criteria and was approved by DESE. The dependent variable for the study was the percent of students scoring proficient and advanced on the MAP in grades three, four, and five for the years 2017, 2018, and 2019. The researcher determined the sum of the percent of students scoring proficient and advanced for each of the evaluation systems, for each year, by grade level for each school district. The quantitative method of research was appropriate for this study as quantitative research is based on relationships between variables (Gay, et al. 2009). Additionally, the findings in quantitative research will typically be generalized to a larger population. The quantitative research design goal is to add to the literature regarding the subject (Gay, et al., 2009).

Data regarding the evaluations used by school districts and MAP results for the years 2017, 2018, and 2019 was collected using DESE's open access MCDS portal. The researcher considered research methods to apply. Experimental research occurs when the researcher's goal is to determine if outcomes will be impacted by a specific treatment. In

experimental research, one group receives the treatment, while the other does not to determine if the outcome was impacted (Creswell, 2009). The design of this study was not to withhold treatment from one group; therefore, the researcher chose not to use an experimental design. The researcher did choose an ex post facto design. Ex post facto was appropriate for this causal-comparative study as ex post facto data is data that has been previously compiled in some manner, but not necessarily for research purposes (Simon & Goes, 2013). In this study, both the evaluations of the teachers and the measures of student achievement have already occurred. When both the cause and the effect have taken place, the research is said to be ex post facto, which is typical of causal-comparative research (Gay, et al., 2009).

Procedures

The researcher retrieved a list of all 518 school districts in the state of Missouri. This list was downloaded into a Microsoft Excel spreadsheet. Using stratified, purposive sampling, the researcher selected specific school districts to utilize in the study. Stratified sampling is used when a researcher wants to ensure the appropriate representatives desired from a subgroup are included in the study (Gay, et al., 2009). The stratified sampling is purposive because the schools selected for the study met the criteria set by the researcher (Gay, et al., 2009). Private and parochial schools were deleted from the list, as they are not required to follow the evaluation requirements. The MCDS web application maintains a list of the evaluation system used by each school district since 2017. The researcher used this list to determine which school districts had used the same evaluation system for the years 2017, 2018, and 2019. Each district that did not use the same evaluation system for all three years was deleted from the list. The list was then

sorted into three groups based on the evaluation system used. Of the 518 school district in Missouri, 86 used the MEES, 240 used the NEE, and 43 used a district created system.

As is required by Southwest Baptist University in regard to research concerning humans, approval for the study was obtained from the Research Review Board (RRB) of the university. The approval needed to complete the study ensures the protection of human participants in a research study. Once approval was obtained from the RRB, data regarding the percent of students scoring proficient and advanced on the MAP was collected from the MCDS web applications on the MODESE open access website for the districts which met the evaluation criteria for the study. The timeline for data collection, analysis, and completion of Chapter Four, and Chapter Five was November 21, 2019 through December 6, 2019. Because data collection was limited to each district's total percentage of students scoring proficient and advanced on the MAP, student and teacher identities were protected, which eliminated the risk of bias or conflict of interest by the researcher. All data collected was maintained by the researcher on a protected Google document, which was only shared with the Southwest Baptist University advisor to further protect the data and identity of each of the school districts.

Following RRB approval, data on MAP performance for each of the school districts selected for the study was collected from the MCDS web application on the MODESE open access website. For each school district, the percent of students scoring proficient and advanced for each grade level for the years 2017, 2018, and 2019 was retrieved. The data was entered into the Microsoft Excel spreadsheet containing the list of school districts, in the corresponding row and column for each school, grade, and year.

Following the data collection and organization into the Excel spreadsheet, an ANOVA was conducted to determine the differences in the scores between the groups.

Instrumentation

Data collected for the study was obtained through the Missouri Department of Elementary and Secondary Education open access website. The Missouri Comprehensive Data System (MCDS) web application at MODESE was used to collect information on the evaluation systems used by each district as well as MAP achievement data in Mathematics and English Language Arts, for the years 2017 through 2019 for grades three through five for each of the districts selected for this study.

In 2015, MODESE began working with the Data Recognition Corporation (DRC) to develop the MAP test, and selected items from their College and Career-Ready items for the MAP (Data Recognition Corporation, 2018). For the first two years after Missouri switched to DRC from Smarter Balanced, the same forms and cut scores were used. Following the 2016-2017 testing cycle, the test was aligned to the revised MLS, new reporting scales had been determined, and cut scores were established for the performance levels. Because of this, MODESE guards against comparing scores from 2017 or prior to the new assessment.

Working with MODESE, DRC seeks to provide valid and reliable results regarding student achievement. According to the Spring 2018 MAP Grade Level Assessment Technical Report, the assessment is evaluated for reliability using Cornbach's coefficient alpha (DRC, 2018). A score between 0 and 1 is obtained, with a score of .8 considered reliable. According to DRC (2018), Cornbach's alpha range is

between .89-.92 for ELA and between .90-.93 for Math, which are both in the acceptable range for reliability (DRC, 2018).

Additionally, DRC provides a variety of question types and methods of scoring to ensure the validity of the assessment (DRC, 2018). A variety of types of questions and requiring different types of responses are provided and the entire assessment is online. For the writing prompts, the DRC provides human readers specifically trained for online assessments. These individuals are recruited, trained, and qualified for evaluating the writing prompts. The trained evaluators are then subject to accuracy checks to ensure scores remain valid. Technology enhanced, evidence-based selected response, and short answer items are scored using DRC's autoscoring engine. DRC has established a set of quality assurance guidelines by which scoring, as well as any rubrics are verified to ensure accuracy and reliability. Multiple-choice and multi-select items are scored during the online administration of the test. For Braille, large-print, or paper-and-pencil versions of the MAP, a test examiner must transcribe the responses into the online system.

Data Analysis

Following the data collection, a one-way ANOVA test was conducted to determine the differences in the percent of third, fourth, and fifth grade students scoring proficient and advanced in 2017, 2018, and 2019 based on the evaluation model used to evaluate the teacher. To run a one-way ANOVA, there are six assumptions that must be taken into consideration (Laerd Statistics, 2016). These six assumptions must be passed for a one-way ANOVA to be used. For this study, the dependent variable was the MAP assessment and the independent variable was the evaluation system used by each of the school districts. The first three assumptions were considered and passed. As an

assessment of academic achievement, the dependent variable, which was the MAP assessment, was measured on a continuous level, which passed the first assumption. The independent variable consisted of three categories of evaluation systems used by the school districts, passed the second assumption of having one independent variable with two or more categorical, independent groups. The third assumption, independence of observation, was also passed, as no participant of any group in this study was a participant of any other group in the study. The data were examined for outliers in the dependent variable, which was the MAP, within the groups of the independent variable, which included the evaluation systems, as required by the fourth assumption to run a one-way ANOVA. Outliers are identified as any data points that lie more than 1.5 box lengths from their own box edge. To accomplish identifying outliers, the researcher used code as described in Laerd Statistics (2016) to create boxplots. Each school district in each subgroup was given a case identifier, which included a unique number. After running the code, any number falling 1.5 boxes, or more, outside the edge of the original box was considered an outlier. To test for normality, as required by assumption five, the Shapiro-Wilk test was conducted on the dependent variable, the MAP, for each group of the independent variables, which included the three evaluation systems. The Shapiro-Wilk test produced a significance level, indicating probability (*p*-value) for each group of the independent variable. This is reported in the "Sig." column. If the data is normally distributed, this value will be greater than .05 (i.e., $p > .05$). If the *p*-value is less than .05 (i.e., $p < .05$), the data is not normally distributed, and the assumption of normality is violated. The final assumption, homogeneity of variances, was tested using Levene's Test of Homogeneity of Variances when the ANOVA was conducted. The test yielded a

significance level (*p*-value), which indicated probability. If the test is not statistically significant, the probability will be greater than .05 (i.e., $p > .05$), indicating homogeneity of variances. However, if the probability is less than .05 (i.e., $p < .05$), the variances are not equal and the assumption of homogeneity of variances has been violated (Laerd Statistics, 2016).

Following the evaluation to meet the assumptions, the one-way ANOVA was conducted to compare the means of the subgroups of the percent of students scoring proficient and advanced in grades three, four, and five on the MAP in 2017, 2018, and 2019 depending on the evaluation system used to evaluate the teacher, to determine any statistically significant differences in the scores. "Simple, or *one-way*, analysis of variance (ANOVA) is a parametric test of significance used to determine whether scores from two or more groups are significantly different at a selected probability level" (Gay, et al., 2016, p. 341). In an ANOVA, two types of variances are considered. The between-groups variance relates to how one group is different from the other groups. The within-groups variance relates to how participants within a particular group differ from each other. To consider these differences, an *F* ratio is determined with the numerator equal to the between-group differences and the numerator equal to the within-group differences. If the between-group variance is larger than the within-group variance, a large ratio would result, which would determine a significant difference. If the within-group variance is larger than the between-group variance, a smaller ratio would result, determining the difference was not significant. In the event that a significant difference is found, the researcher will know the groups are not the same, but cannot determine specifically how

the groups differ (Gay, et al., 2009). For this study, the researcher was most interested in identifying the between groups variance.

Because the independent variable contains three subgroups of the evaluation systems used, more than one test of significance must be conducted on the same dataset, which increases the chance for error. Additionally, the ANOVA will indicate if there are significant differences, but it will not tell where the differences are (Gay, et al., 2009). Because there are three subgroups of the evaluation systems used, the one-way ANOVA will be run with the Tukey/Honest Significant Difference post hoc multiple comparison test. The test will compare every pairing of the calculated means of the subgroups to determine which group's means differ (Laerd Statistics, 2016). After identifying whether there was a statistically significant difference, and which subgroup of the data were statistically different, the effect size of the difference was determined. Determining effect size was conducted with a general linear model which calculated the total sum of squares and divided it into the between groups sum of squares (Laerd Statistics, 2016). The researcher was aiming for a medium effect size for this study.

In this study, the researcher sought to determine if there was a statistically significant difference in the percent of students scoring proficient and advanced on the MAP among school districts using one of three evaluation models approved by MODESE for use in Missouri school districts. The null hypothesis for the comparison of the means was there would be no statistical difference between the means of the percent of students scoring proficient and advanced on the MAP among school districts using one of three evaluation models approved by MODESE for use in Missouri school districts. The

researcher compared total populations of students from each district and did not disaggregate data into demographic groups.

To determine statistical significance, alpha level must be set prior to completing a study. Alpha is the point at which it is determined an outcome cannot be attributed to chance (Bruce, 2015). If an outcome occurs frequently, the researcher cannot rule out chance as the possible cause. Alpha is typically set at 0.05, or five percent, based on the work of statistician R. A. Fisher close to 100 years ago. Fisher determined that an occurrence of once in 20 trials was rare enough to determine the outcome was likely not a result of chance and therefore statistically significant (Bruce, 2015). For the purpose of this study, an alpha level of five percent was used. Additionally, the researcher sought a medium effect size for this study. According to Gay, et al. (2009), effect size represents the strength of the relationship between the variables in the study. Effect size can have a negative or positive value and is expressed a decimal. An effect size around .20 would indicate a small effect while an effect size around .80 would indicate a large effect (Gay, et al., 2009). Effect size is calculated once the difference in the percent of students scoring proficient and advanced for each subgroup for each grade level in each year has been determined.

Summary

The purpose of this quantitative, causal-comparative study was to determine the differences in the percent of students scoring proficient and advanced on the MAP, depending on the evaluation system used to evaluate the teacher for elementary students in grades three, four, and five in Missouri between the years 2017 and 2019. Data on the evaluation used by each school district in Missouri was collected from MODESE's

MCDS portal. Stratified, purposive sampling was chosen to select the districts for this study. Each Missouri school district using one of the evaluation systems for the years 2017, 2018, and 2019 was chosen for the study. Once identified, the percent of students scoring proficient and advanced on the MAP for students in third, fourth, and fifth grade was collected for each of the school districts selected.

Analysis of the data was conducted using a one-way ANOVA with a Tukey/Honest Significant Difference post hoc multiple comparison test to compare the means. The test was conducted to determine if there were differences in the percent of students scoring proficient and advanced, depending on which evaluation system was used to evaluate the teacher. The effect size was calculated to determine if a minimum medium effect size was achieved.

Chapter Four will present an analysis of the data and the results of the research study. The results will be explained in relation to the research questions posed in the study. The null hypotheses as related to the results of the data analyses will be addressed as well.

Chapter Four

Analysis of the Data

Introduction

The purpose of this quantitative, causal-comparative study was to determine the differences in the percent of students scoring proficient and advanced on the Missouri Assessment Program (MAP) in third, fourth and fifth grades, depending on which evaluation system was used to evaluate the teacher between the years 2017 and 2019. The independent variable was the evaluation system used among the schools. The choices of the evaluation systems included the Missouri Educator Evaluation System (MEES), the Network for Educator Effectiveness (NEE), or a district created evaluation system. The dependent variable was the sum of the proficient and advanced scores on the Missouri Assessment Program (MAP) scores for the years 2017 through 2019. The researcher explored the differences in the percent of students scoring proficient and advanced on the MAP among schools using one of three evaluation models approved by the Missouri Department of Elementary and Secondary Education (MODESE).

Improved teacher evaluation has been an important topic for states since federal legislation has focused on improving evaluation measures and student achievement. No Child Left Behind (NCLB), Race to the Top (RTTT), and the Every Student Succeeds Act (ESSA) all focused on increasing student achievement and holding teachers accountable (Donaldson, 2016). With recent legislation in mind, the researcher chose to look at school districts that used one of three evaluation systems approved by MODESE; the MEES, the evaluation model created by MODESE, the NEE, the evaluation model created by Missouri University, and district created models approved by MODESE for

use in the school district. Each of the school districts chosen for the study used the same evaluation model for the years 2017, 2018, and 2019. The researcher then retrieved proficient and advanced percentage MAP results for third, fourth, and fifth grade, for each of the school districts for the years 2017, 2018, and 2019. Using this data, a one-way Analysis of Variance (ANOVA) was conducted to find the differences, and whether there were statistically significant differences in the data among schools using the three evaluation systems. To accomplish this task, the researcher developed a set of research questions. In this chapter, the research questions and null hypotheses are shared and an analysis and results of the data are presented.

Research Questions. The following research questions have been developed to address the stated problem:

1. What are the differences in the **third grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system and DESE approved, district developed evaluation model in **2017**?
2. What are the differences in the **fourth grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Missouri Educator Evaluation System, the Network for Educator Effectiveness evaluation system and DESE approved, district developed evaluation model in **2017**?
3. What are the differences in the **fifth grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the

Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system and DESE approved, district developed evaluation model in **2017**?

4. What are the differences in the **third grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system and DESE approved, district developed evaluation model in **2018**?
5. What are the differences in the **fourth grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system and DESE approved, district developed evaluation model in **2018**?
6. What are the differences in the **fifth grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system and DESE approved, district developed evaluation model in **2018**?
7. What are the differences in the **third grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system and DESE approved, district developed evaluation model in **2019**?

8. What are the differences in the **fourth grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system and DESE approved, district developed evaluation model in **2019**?
9. What are the differences in the **fifth grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system and DESE approved, district developed evaluation model in **2019**?

Null Hypotheses. The following null hypotheses were investigated to answer the research questions:

$H_0$1. There are no statistically significant differences in the **third grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system and a DESE approved, district developed evaluation model in **2017**.

$H_0$2. There are no statistically significant differences in the **fourth grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system and a DESE approved, district developed evaluation model in **2017**.

$H_0$3. There are no statistically significant differences in the **fifth grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system and a DESE approved, district developed evaluation model in **2017**.

$H_0$4. There are no statistically significant differences in the **third grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system and a DESE approved, district developed evaluation model in **2018**.

$H_0$5. There are no statistically significant differences in the **fourth grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system and a DESE approved, district developed evaluation model in **2018**.

$H_0$6. There are no statistically significant differences in the **fifth grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system and a DESE approved, district developed evaluation model in **2018**.

$H_0$7. There are no statistically significant differences in the **third grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator

Effectiveness evaluation system and a DESE approved, district developed evaluation model in **2019**.

$H_0$8. There are no statistically significant differences in the **fourth grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system and a DESE approved, district developed evaluation model in **2019**.

$H_0$9. There are no statistically significant differences in the **fifth grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system and a DESE approved, district developed evaluation model in **2019**.

In Chapter Four, an analysis of the data and findings are shared. The samples tested and their demographics are presented along with the data cleaning process. For each null hypothesis, results of the assumptions testing are presented, followed by the research findings, a summary of the data analysis, and an interpretation of the outcomes. Finally, the researcher will either accept or reject the null hypotheses, based on the statistical conclusion reached. The following is the analysis of the data and the findings of the research study conducted.

Data Analysis

This study explored the differences in student achievement of students in third, fourth, and fifth grade public school districts in Missouri for the years 2017, 2018, and 2019 among schools using the MEES, NEE, and a district created evaluation system. The

independent variables were the evaluation systems used, and the dependent variables were the sum of the proficient and advanced MAP scores. The MCDS portal on MODESE Web applications contains information on which evaluation system each school district uses each year. The information was retrieved, and a list of schools was created in a Microsoft Excel document. Schools that did not use the same evaluation system, the MEES, NEE, or a district created evaluation system for the years 2017, 2018, and 2019 were removed from the list.

After determining the list of schools to include in the sample, the percent of students scoring proficient and advanced was retrieved from MODESE's MCDS portal for English language arts and math for each of the three years. The percentages were entered into the Microsoft Excel spreadsheet by year. School districts that did not have qualifying scores in the proficient and advanced categories were removed from the list. School districts with no data were also removed from the list. School districts that did not have data for all three years, in all three grade levels were also removed from the list. The remaining data was used to run the one-way ANOVA to compare the means for each grade level for the purpose of identifying any statistically significant differences in the percent of students scoring proficient and advanced among the three evaluation systems, the MEES, the NEE, and a district created evaluation system.

Samples. The total number of samples collected for school districts using the MEES in 2017, 2018, and 2019 for grades three, four, and five was 177. Of those samples, 67 were missing, which left 110 valid samples. The total number of samples collected for school districts using the NEE in 2017, 2018, and 2019 for grades three, four, and five was 606. Of those samples, 176 were missing, which left 430 valid

samples. The total number of samples collected for school districts using a district created evaluation in 2017, 2018, and 2019 for grades three, four, and five was 111. Of those samples, 28 were missing, which left 73 valid samples; see Table 1.

Table 1

Case Processing Summary

		Cases					
		Valid		Missing		Total	
	Eval	N	Percent	N	Percent	N	Percent
ELA2019	MEES	110	62.1%	67	37.9%	177	100.0%
	NEE	430	71.0%	176	29.0%	606	100.0%
	District	83	74.8%	28	25.2%	111	100.0%
ELA2018	MEES	110	62.1%	67	37.9%	177	100.0%
	NEE	430	71.0%	176	29.0%	606	100.0%
	District	83	74.8%	28	25.2%	111	100.0%
ELA2017	MEES	110	62.1%	67	37.9%	177	100.0%
	NEE	430	71.0%	176	29.0%	606	100.0%
	District	83	74.8%	28	25.2%	111	100.0%
Math2019	MEES	110	62.1%	67	37.9%	177	100.0%
	NEE	430	71.0%	176	29.0%	606	100.0%
	District	83	74.8%	28	25.2%	111	100.0%
Math2018	MEES	110	62.1%	67	37.9%	177	100.0%
	NEE	430	71.0%	176	29.0%	606	100.0%
	District	83	74.8%	28	25.2%	111	100.0%
Math2017	MEES	110	62.1%	67	37.9%	177	100.0%
	NEE	430	71.0%	176	29.0%	606	100.0%
	District	83	74.8%	28	25.2%	111	100.0%

Demographics. Third, fourth, and fifth grade MAP results from school districts in the state of Missouri were used for this study. The teachers had an average of 12.8 years of experience in 2017, 12.8 years of experience in 2018, and 12.9 percent in 2019. Average class sizes for the years of the study were 17 in 2017, 2018, and 2019. The average student to principal ratio for the years of the study was 183 in 2017, 181 in 2018, and 178 in 2019. The attendance rate for each year was 88.7 percent in 2017, 87.7 percent

in 2018, and 87.3 percent in 2019. Males and females typically between the ages of eight and nine were included the third-grade results. Males and females typically between the ages of nine and ten were included the fourth-grade results. Males and females typically between the ages of ten and eleven were included in the fifth-grade results.

Data Cleaning. Accurate results are essential in a research study; therefore, it is important to look for errors and inaccuracies in the data collected. The first data-cleaning task dealt with missing data. Missing data, or those with a value of zero, were removed from the sample to prevent the data from affecting the outcome of the ANOVA. This was acceptable because the sample collected was large enough to run the ANOVA (Laerd Statistics, 2016). After dealing with the missing data, the data was examined for the last three assumptions of the one-way ANOVA. The first three assumptions were met and addressed in Chapter Three. As an assessment of academic achievement, the dependent variable, the MAP assessment, was measured on a continuous level and passed the first assumption. The independent variable consisted of three categories of evaluation systems used by the school districts and passed the second assumption of having one independent variable with two or more categorical, independent groups. The third assumption, independence of observation, was passed, as no participant of any group in the study was a participant of any other group in the study.

The fourth assumption states the data should have no significant outliers in the groups of the independent variables, for each of the dependent variable (Laerd Statistics, 2016). The fourth assumption was determined by running the Explore procedure in SPSS. Boxplots were created for each year of the dependent variable for both ELA and Math (See Appendix A-F). There were outliers in the data, as assessed by inspection of a

boxplot for values greater than 1.5 times the interquartile range above the third quartile or below the first quartile. The outliers found were of two types. The first were data entry errors, which were located and corrected using the value from the original data retrieved from MODESE's MCDS portal. The second type of error was a report of 100 percent of students scoring proficient or advanced on the MAP. These data entries were verified as correct. A report of 100 percent occurred rarely and was included in the analysis because they were not believed to have a material effect on the result (Laerd Statistics, 2016). ELA 2019, ELA 2017, math 2019, math 2018, and math 2017 proficient and advanced percentages were normally distributed for the MEES, NEE, and district evaluation systems, as assessed by the Shapiro-Wilk's test ($p > .05$). ELA 2018 proficient and advanced percentages were normally distributed for the district evaluation system, as assessed by the Shapiro-Wilk's test ($p > .05$). The null hypothesis for the Shapiro-Wilk's test, which indicates the data's distribution is equal to a normal distribution, is accepted for these samples. ELA 2018 proficient and advanced percentages were not normally distributed ($p < .05$) for the MEES or NEE as assessed by the Shapiro-Wilk's test (see Table 2). Therefore, the Shapiro-Wilk test null hypothesis is rejected for ELA 2018 for the MEES and NEE evaluations and the alternative hypothesis is accepted.

Table 2
Tests of Normality

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
MAP	Eval	Statistic	df	Sig.	Statistic	df	Sig.
ELA2019	MEES	.083	110	.061	.980	110	.101
	NEE	.004	430	.043	.994	430	.066
	District	.082	83	.200*	.981	83	.264
ELA2018	MEES	.104	110	.005	.963	110	.004
	NEE	.127	430	.000	.585	430	.000
	District	.067	83	.200*	.984	83	.393
ELA2017	MEES	.060	110	.200*	.983	110	.173
	NEE	.027	430	.200*	.996	430	.283
	District	.058	83	.200*	.981	83	.274
Math2019	MEES	.050	110	.200*	.991	110	.644
	NEE	.032	430	.200*	.996	430	.404
	District	.083	83	.200*	.980	83	.234
Math2018	MEES	.087	110	.041	.979	110	.080
	NEE	.041	430	.080	.996	430	.438
	District	.066	83	.200*	.982	83	.288
Math2017	MEES	.063	110	.200*	.980	110	.094
	NEE	.034	430	.200*	.996	430	.417
	District	.071	83	.200*	.981	83	.253

*This is a lower bound of the true significance.
a. Lilliefors Significance Correction

While the null hypothesis could not be accepted for two of the samples, the researcher chose to proceed with running the one-way ANOVA. The one-way ANOVA can provide acceptable results, as ANOVA is able to handle deviations from normal sample size distributions (Laerd Statistics, 2016). The final assumption, the test for homogeneity of variances, was conducted when the ANOVA was run and is presented with the findings.

Results

A one-way ANOVA was conducted to compare the means to find the differences in the students scoring proficient and advanced for the years 2017, 2018, and 2019 in grades three, four, and five among the evaluation systems used. In this section, the results

are presented with each research question and null hypothesis. The results of the homogeneity of variances are presented along with a summary of the analysis.

Research Question 1: What are the differences in the **third grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system and DESE approved, district developed evaluation model in **2017**?

$H_0$1. There are no statistically significant differences in the **third grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system and a DESE approved, district developed evaluation model in **2017**.

Table 3

Welch ANOVA Third Grade 2017

	F	*df1*	*df2*	*Sig.*
ELA2017	2.299	2	59.023	.109
Math2017	1.966	2	60.463	.149

Table 4

Tukey Multiple Comparisons Third Grade 2017

Dependent Variable	(I) Eval	(J) Eval	Mean Difference (I-J)	Std. Error	*Sig.*	95% Confidence Interval Lower Bound	Upper Bound
ELA2017	MEES	NEE	-4.518	2.287	.120	-9.91	.87
		District	-7.638	3.233	.049	-15.26	-.01
	NEE	MEES	4.518	2.287	.120	-.87	9.91
		District	-3.120	2.726	.488	-9.55	3.31
	District	MEES	7.638	3.233	.049	.01	15.26
		NEE	3.120	2.726	.488	-3.31	9.55
Math2017	MEES	NEE	-5.188	2.632	.122	-11.40	1.02
		District	-7.878	3.716	.088	-16.64	.89
	NEE	MEES	5.188	2.632	.122	-1.02	11.4
		District	-2.689	3.136	.688	-10.08	4.71
	District	MEES	7.878	3.716	.088	-.89	16.64
		NEE	2.689	3.136	.668	-4.71	10.08

A one-way Welch ANOVA was conducted to determine the differences in the third grade proficient and advanced percentages of the 2017 ELA MAP scores among school districts using the three different evaluation systems (see Table 3). A Welch ANOVA is used to compare the means when the assumption of homogeneity of variances is violated (Laerd Statistics, 2016). School districts were classified into three groups: MEES (n = 46), NEE (*n* = 172), and District (*n* = 30). There were no outliers as assessed by boxplot (see Appendix A); data was normally distributed for each group, as assessed by Shapiro-Wilk test ($p > .05$); but there was heterogeneity of variances, as assessed by Levene's test of homogeneity of variances ($p = .041$). The 2017 third grade ELA MAP proficient and advanced percentages increased from the MEES ($M = 56.3$, $SD = 16.8$), to NEE ($M = 60.8$, $SD = 12.7$), to District ($M = 64.0$, $SD = 14.4$), in the said order, but the differences between the evaluation systems was not statistically significant, Welch's $F(2, 59.023) = 2.299$, $p = .109$ (see Table 3). The group means were not statistically significantly different ($p = .109$). Therefore, the null hypothesis failed to be rejected.

A one-way Welch ANOVA was conducted to determine the differences in the third grade proficient and advanced percentages of the 2017 Math MAP scores among school districts using the three different evaluation systems (see Table 3). School districts were classified into three groups: MEES (n = 46), NEE (*n* = 170), and District (*n* = 30). There were no outliers as assessed by boxplot (see Appendix D); data was normally distributed for each group, as assessed by Shapiro-Wilk test ($p > .05$); but there was heterogeneity of variances, as assessed by Levene's test of homogeneity of variances ($p = .003$). The 2017 third grade math MAP proficient and advanced percentages increased from the MEES ($M = 48.2$, $SD = 20.2$), to NEE ($M = 53.4$, $SD = 14.6$), to District ($M =$

56.1, *SD* = 14.9), in the said order, but the differences between the evaluation systems was not statistically significant, Welch's $F(2, 60.463) = 1.966$, $p = .149$ (see Table 3). The group means were not statistically significantly different ($p = .149$). Therefore, the null hypothesis failed to be rejected.

Research Question 2: What are the differences in the **fourth grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Missouri Educator Evaluation System, the Network for Educator Effectiveness evaluation system, and DESE approved, district developed evaluation model in **2017**? $H_0$2. There are no statistically significant differences in the **fourth grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system and a DESE approved, district developed evaluation model in **2017**.

Table 5

ANOVA Fourth Grade 2017

		Sum of Squares	*df*	Mean Square	*F*	*Sig.*
ELA	Between Groups	1844.136	2	922.068	5.407	.005
	Within Groups	41779.396	245	170.528		
	Total	43623.532	247			
Math	Between Groups	1525.771	2	762.885	3.473	.033
	Within Groups	53602.787	244	219.684		
	Total	55128.558	246			

Table 6

Tukey Multiple Comparisons Fourth Grade 2017

Dependent Variable	(I) Eval	(J) Eval	Mean Difference (I-J)	Std. Error	*Sig.*	95% Confidence Interval Lower Bound	Upper Bound
ELA2017	MEES	NEE	-7.069*	2.151	.003	-12.14	-2.0
		District	-5.813	3.052	.139	-13.01	1.38
	NEE	MEES	7.069*	2.151	.003	2.00	12.14
		District	1.256	2.585	.878	-4.84	7.35
	District	MEES	5.813	3.052	.139	-1.38	13.01
		NEE	-1.256	2.585	.878	-7.35	4.84
Math2017	MEES	NEE	-6.486*	2.462	.024	-12.29	-.68
		District	-4.917	3.478	.335	-13.12	3.29
	NEE	MEES	6.486*	2.462	.024	.68	12.29
		District	1.569	2.934	.854	-5.35	8.49
	District	MEES	4.917	3.478	.335	-3.29	13.12
		NEE	-1.569	2.934	.854	-8.49	5.35

*. The mean difference is significant at the 0.05 level.

A one-way ANOVA was conducted to determine the differences in the fourth grade proficient and advanced percentages of the 2017 ELA MAP scores among school districts that use the three different evaluations systems (see Table 5). School districts were classified into three groups: MEES (n = 47), NEE (n = 171), and District (n = 30). There were no outliers, as assessed by boxplot (see Appendix A); data was normally distributed for each group, as assessed by Shapiro-Wilk test ($p > .05$); and there was homogeneity of variances, as assessed by Levene's test of homogeneity of variances (p = .857). Data was presented as a mean ± standard deviation. The differences in the fourth grade proficient and advanced percentages of the 2017 ELA MAP scores were statistically significantly different between different evaluations, $F(2, 245) = 5.407$, p = .005. The differences in the fourth grade proficient and advanced percentages of the 2017 ELA MAP scores among school districts for the three different evaluations increased from MEES (M = 57.1, SD = 14.0), to District (M = 63.0, SD = 12.7), to NEE (M = 64.2,

SD = 12.9), in the said order. In Table 6, Tukey's post hoc analysis revealed the mean increase from the MEES to the NEE (7.069, 95% CI [2.00, 12.14]) was statistically significant at (p = .003). A small effect size was calculated using Cohen's d (d = .11). A small effect size indicates a weak relationship between the variables (McLeod, 2019). No other group differences were statistically significant. The group means were statistically significantly different (p = .005) according to the ANOVA. Therefore, the null hypothesis was rejected.

A one-way ANOVA was conducted to determine the differences in the fourth grade proficient and advanced percentages of the 2017 math MAP scores among school districts that use the three different evaluations systems (see Table 5). School districts were classified into three groups: MEES (n = 46), NEE (n = 171), and District (n = 30). There were no outliers, as assessed by boxplot (see Appendix D); data was normally distributed for each group, as assessed by Shapiro-Wilk test (p > .05); and there was homogeneity of variances, as assessed by Levene's test of homogeneity of variances (p = .995). Data was presented as a mean ± standard deviation. The differences in the fourth grade proficient and advanced percentages of the 2017 math MAP scores was statistically significantly different between different evaluations, $F(2, 244) = 3.473$, p = .033. The differences in the fourth grade proficient and advanced percentages of the 2017 math MAP scores among school districts for the three different evaluations increased from MEES (M = 47.22, SD = 14.7), to District (M = 52.1, SD = 15.1), to NEE (M = 53.7, SD = 14.8), in the said order. In Table 6, Tukey post hoc analysis revealed the mean increase from the MEES to the NEE (6.486, 95% CI [.68, 12.29]) was statistically significant at (p = .024). A small effect size was calculated using Cohen's d (d = .12). A small effect size

indicates a weak relationship between the variables (McLeod, 2019). No other group differences were statistically significant. The group means were statistically significantly different (p = .033) according to the ANOVA. Therefore, the null hypothesis was rejected.

Research Question 3: What are the differences in the **fifth grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system and DESE approved, district developed evaluation model in **2017**?

$H_0$3. There are no statistically significant differences in the **fifth grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system and a DESE approved, district developed evaluation model in **2017**.

Table 7

ANOVA Fifth Grade 2017

		Sum of Squares	*df*	Mean Square	*F*	*Sig.*
ELA	Between Groups	1309.227	2	654.613	3.757	.025
	Within Groups	42868.049	246	174.260		
	Total	44177.276	248			
Math	Between Groups	2698.536	2	1349.268	5.275	.006
	Within Groups	60105.105	235	255.766		
	Total	62803.641	237			

Table 8

Tukey Multiple Comparisons Fifth Grade 2017

Dependent Variable	(I) Eval	(J) Eval	Mean Difference (I-J)	Std. Error	*Sig.*	95% Confidence Interval Lower Bound	Upper Bound
ELA2017	MEES	NEE	-5.936*	2.173	.018	-11.06	-.81
		District	-5.232	3.085	.209	-12.51	2.04
	NEE	MEES	5.936*	2.173	.018	.81	11.06
		District	.704	2.612	.961	-5.45	6.86
	District	MEES	5.232	3.085	.209	-2.04	12.51
		NEE	-.704	2.612	.961	-6.86	5.45
Math2017	MEES	NEE	-8.883*	2.735	.004	-15.33	-2.43
		District	-7.068	3.884	.165	-16.23	2.09
	NEE	MEES	8.883*	2.735	.004	2.43	15.33
		District	1.815	3.226	.844	-5.89	9.52
	District	MEES	7.068	3.884	.165	-2.09	16.23
		NEE	-1.815	3.226	.844	-9.52	5.89

*. The mean difference is significant at the 0.05 level.

A one-way ANOVA was conducted to determine the differences in the fifth grade proficient and advanced percentages of the 2017 ELA MAP scores among school districts that use the three different evaluations systems (see Table 7). School districts were classified into three groups: MEES (n = 47), NEE (n = 172), and District (n = 30). There were no outliers, as assessed by boxplot (see Appendix A); data was normally distributed for each group, as assessed by Shapiro-Wilk test ($p > .05$); and there was homogeneity of variances, as assessed by Levene's test of homogeneity of variances ($p = .146$). Data was presented as a mean ± standard deviation. The differences in the fifth grade proficient and advanced percentages of the 2017 ELA MAP scores was statistically significantly different between different evaluations, $F(2, 246) = 3.757$, $p = .025$. The differences in the fifth grade proficient and advanced percentages of the 2017 ELA MAP scores among school districts for the three different evaluations increased from MEES (M = 56.0, SD = 14.6), to District (M = 61.2, SD = 15.2), to NEE (M = 61.9, SD = 12.4), in the said order.

In Table 8, Tukey's post hoc analysis revealed the mean increase from the MEES to the NEE (5.936, 95% CI [.81, 11.06]) was statistically significant at (p = .018). A small effect size was calculated using Cohen's d (d = .10). A small effect size indicates a weak relationship between the variables (McLeod, 2019). No other group differences were statistically significant. The group means were statistically significantly different (p = .025) according to the ANOVA. Therefore, the null hypothesis was rejected.

A one-way ANOVA was conducted to determine the differences in the fifth grade proficient and advanced percentages of the 2017 math MAP scores among school districts that use the three different evaluations systems (see Table 7). School districts were classified into three groups: MEES (n = 43), NEE (n = 167), and District (n = 28). There were no outliers, as assessed by boxplot (see Appendix D); data was normally distributed for each group, as assessed by Shapiro-Wilk test (p > .05); and there was homogeneity of variances, as assessed by Levene's test of homogeneity of variances (p = .585). Data was presented as a mean ± standard deviation. The differences in the fifth grade proficient and advanced percentages of the 2017 math MAP scores was statistically significantly different between different evaluations, $F(2, 235) = 5.275$, p = .006. The differences in the fifth grade proficient and advanced percentages of the 2017 Math MAP scores among school districts for the three different evaluations increased from MEES (M = 39.8, SD = 15.7), to NEE (M = 48.7, SD = 15.8), to District (M = 46.9, SD = 17.7), in the said order. In Table 8, Tukey's post hoc analysis revealed the mean increase from the MEES to the NEE (8.883, 95% CI [2.43, 15.33]) was statistically significant at (p = .004). A small effect size was calculated using Cohen's d (d = .19). A small effect size indicates a weak relationship between the variables (McLeod, 2019). No other group

differences were statistically significant. The group means were statistically significantly different (p = .006) according to the ANOVA. Therefore, the null hypothesis was rejected.

Research Question 4: What are the differences in the **third grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system, and DESE approved, district developed evaluation model in **2018**?

$H_0$4. There are no statistically significant differences in the **third grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system and a DESE approved, district developed evaluation model in **2018**.

Table 9

ANOVA Third Grade 2018

		Sum of Squares	*df*	Mean Square	*F*	*Sig.*
ELA	Between Groups	598.274	2	299.136	1.663	.191
	Within Groups	52689.915	293	179.829		
	Total	43288.188	295			
Math	Between Groups	300.952	2	150.476	.721	.487
	Within Groups	55936.627	268	208.719		
	Total	56237.579	270			

Table 10

Tukey Multiple Comparisons Third Grade 2018

Dependent Variable	(I) Eval	(J) Eval	Mean Difference (I-J)	Std. Error	*Sig.*	95% Confidence Interval Lower Bound	Upper Bound
ELA2018	MEES	NEE	-1.583	1.999	.708	-6.29	3.13
		District	-5.101	2.821	.169	-11.75	1.55
	NEE	MEES	1.583	1.999	.708	-3.13	6.29
		District	-3.518	2.399	.309	-9.17	2.13
	District	MEES	5.101	2.821	.169	-1.55	11.75
		NEE	3.518	2.399	.309	-2.13	9.17
Math2018	MEES	NEE	-3.300	2.251	.309	-8.61	2.01
		District	-5.652	3.144	.172	-13.06	1.76
	NEE	MEES	3.300	2.251	.309	-2.01	8.61
		District	-2.352	2.663	.651	-8.63	3.92
	District	MEES	5.652	3.144	.172	-1.76	13.06
		NEE	2.352	2.663	.651	-3.92	8.63

A one-way ANOVA was conducted to determine the differences in the third grade proficient and advanced percentages of the 2018 ELA MAP scores among school districts that use the three different evaluation systems (see Table 9). School districts were classified into three groups: MEES (n = 58), NEE (n = 201), and District (n = 207). There were no outliers, as assessed by boxplot (see Appendix B); data was normally distributed for each group, as assessed by Shapiro-Wilk test ($p > .05$); and there was homogeneity of variances, as assessed by Levene's test of homogeneity of variances (p = .569). The 2018 third grade ELA MAP proficient and advanced percentages increased from the MEES (M = 45.9, SD = 14.4), to NEE (M = 47.5, SD = 12.0), to District (M = 51.0, SD = 14.2), in the said order, but the differences between the evaluation systems was not statistically significant, $F(2, 293) = 1.663$, p = .191. The group means were not statistically significantly different (p = .191). Therefore, the null hypothesis failed to be rejected.

A one-way ANOVA was conducted to determine the differences in the third grade proficient and advanced percentages of the 2018 math MAP scores among school districts that use the three different evaluation systems (see Table 9). School districts were classified into three groups: MEES (n = 55), NEE (n = 180), and District (n = 36). There were no outliers, as assessed by boxplot (see Appendix E); data was normally distributed for each group, as assessed by Shapiro-Wilk test ($p > .05$); and there was homogeneity of variances, as assessed by Levene's test of homogeneity of variances (p = .542). The 2018 third grade math MAP proficient and advanced percentages increased from the MEES (M = 45.0, SD = 15.0), to NEE (M = 46.8, SD = 13.0), to District (M = 48.7, SD = 14.3), in the said order, but the differences between the evaluation systems was not statistically significant, $F(2, 268) = .721$, p = .487. The group means were not statistically significantly different (p = .487). Therefore, the null hypothesis failed to be rejected.

Research Question 5: What are the differences in the **fourth grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system and DESE approved, district developed evaluation model in **2018**?

$H_0$5. There are no statistically significant differences in the **fourth grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system and a DESE approved, district developed evaluation model in **2018**.

Table 11

ANOVA Fourth Grade 2018

		Sum of Squares	*df*	Mean Square	*F*	*Sig.*
ELA	Between Groups	2048.727	2	1024.364	5.803	.003
	Within Groups	51370.096	291	176.530		
	Total	53418.824	293			
Math	Between Groups	188.405	2	94.203	.419	.658
	Within Groups	59646.373	265	225.081		
	Total	59834.779	267			

Table 12

Tukey Multiple Comparisons Fourth Grade 2018

Dependent Variable	(I) Eval	(J) Eval	Mean Difference (I-J)	Std. Error	*Sig.*	95% Confidence Interval Lower Bound	Upper Bound
ELA2018	MEES	NEE	-6.581*	1.995	.003	-11.28	-1.88
		District	-7.106*	2.805	.032	-13.71	-.50
	NEE	MEES	6.581*	1.995	.003	1.88	11.28
		District	-525	2.378	.974	-6.13	5.08
	District	MEES	7.106*	2.805	.032	.50	13.71
		NEE	.525	2.378	.974	-5.08	6.13
Math2018	MEES	NEE	-1.909	2.329	.691	-7.40	3.58
		District	-2.573	3.256	.709	-10.25	5.10
	NEE	MEES	1.909	2.329	.691	-3.58	7.40
		District	-.665	2.773	.969	-7.20	5.87
	District	MEES	2.573	3.256	.709	-5.10	10.25
		NEE	.665	2.773	969	-5.87	7.20

A one-way ANOVA was conducted to determine the differences in the fourth grade proficient and advanced percentages of the 2018 ELA MAP scores among school districts that use the three different evaluations systems (see Table 11). School districts were classified into three groups: MEES (n = 57), NEE (n = 200), and District (n = 37). There were no outliers, as assessed by boxplot (see Appendix B); data was normally distributed for District (p = .393), but was not normally distributed for MEES (p = .004), or NEE (p = .000), as assessed by Shapiro-Wilk test; and there was homogeneity of variances, as assessed by Levene's test of homogeneity of variances (p = .752). Data was

presented as a mean ± standard deviation. The differences in the fourth grade proficient and advanced percentages of the 2018 ELA MAP scores was statistically significantly different between different evaluations, $F(2, 291) = 5.803$, $p = .003$. Differences in fourth grade proficient and advanced percentages for the 2018 ELA MAP among school districts for the three different evaluations increased from MEES ($M = 57.1$, $SD = 14.0$), to District ($M = 63.0$, $SD = 12.7$), to NEE ($M = 64.2$, $SD = 12.9$), in the said order. In Table 12, Tukey' post hoc analysis revealed the mean increase from the MEES to the NEE (6.581, 95% CI [1.88, 11.28]) was statistically significant ($p = .003$), with a small effect size (d = .14), as well as from the MEES to the District (7.106, 95% CI [.50, 13.71], $p = .032$). A small effect size was calculated using Cohen's d ($d = .15$). A small effect size indicates a weak relationship between the variables (McLeod, 2019). No other group differences were statistically significant. The group means were statistically significantly different ($p = .003$) according to the ANOVA. Therefore, the null hypothesis was rejected.

A one-way ANOVA was conducted to determine the differences in the fourth grade proficient and advanced percentages of the 2018 math MAP scores among school districts that use the three different evaluation systems (see Table 11). School districts were classified into three groups: MEES ($n = 54$), NEE ($n = 179$), and District ($n = 35$). There were no outliers, as assessed by boxplot (see Appendix E); data was normally distributed for each group, as assessed by Shapiro-Wilk test ($p > .05$); and there was homogeneity of variances, as assessed by Levene's test of homogeneity of variances ($p = .159$). The 2018 fourth grade math MAP proficient and advanced percentages increased from the MEES ($M = 43.5$, $SD = 17.5$), to NEE ($M = 45.4$, $SD = 14.3$), to District ($M =$

46.1, *SD* = 14.5), in the said order, but the differences between the evaluation systems was not statistically significant, $F(2, 265) = .419$, $p = .658$. The group means were not statistically significantly different ($p = .658$). Therefore, the null hypothesis failed to be rejected.

Research Question 6: What are the differences in the **fifth grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system and DESE approved, district developed evaluation model in **2018**?

$H_0$6. There are no statistically significant differences in the **fifth grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system and a DESE approved, district developed evaluation model in **2018**.

Table 13

ANOVA Fifth Grade 2018

		Sum of Squares	*df*	Mean Square	*F*	*Sig.*
ELA	Between Groups	977.831	2	488.916	1.133	.324
	Within Groups	125621.706	291	431.690		
	Total	126599.538	293			

Table 14

Tukey Multiple Comparisons Fifth Grade 2018

Dependent Variable	(I) Eval	(J) Eval	Mean Difference (I-J)	Std. Error	*Sig.*	95% Confidence Interval Lower Bound	Upper Bound
ELA2018	MEES	NEE	-4.692	3.120	.290	-12.04	2.66
		District	-3.459	4.386	.710	-13.79	6.87
	NEE	MEES	4.692	3.120	.290	-2.66	12.04
		District	1.233	3.718	.941	-7.53	9.99
	District	MEES	3.459	4.386	.710	-6.87	13.79
		NEE	-1.233	3.718	.941	-9.99	7.53
Math2018	MEES	NEE	-2.174	2.383	.633	-7.79	3.44
		District	-1.281	3.283	.920	-9.02	6.46
	NEE	MEES	2.174	2.383	.633	-3.44	7.79
		District	.894	2.782	.945	-5.66	7.45
	District	MEES	1.281	3.283	.920	-6.46	9.02
		NEE	-.894	2.782	.945	-7.45	5.66

A one-way ANOVA was conducted to determine the differences in the fifth grade proficient and advanced percentages of the 2018 ELA MAP scores among school districts that use the three different evaluation systems (see Table 13). School districts were classified into three groups: MEES (n = 57), NEE (n = 200), and District (n = 37). There were no outliers, as assessed by boxplot (see Appendix B); data was normally distributed for each group, as assessed by Shapiro-Wilk test ($p > .05$); and there was homogeneity of variances, as assessed by Levene's test of homogeneity of variances ($p = .738$). The 2018 fifth grade ELA MAP proficient and advanced percentages increased from the MEES (M = 42.2, SD = 14.1), to District (M = 45.7, SD = 11.2), to NEE (M = 46.9, SD = 23.5), in the said order, but the differences between these evaluation systems was not statistically significant, $F(2, 291) = 1.133$, $p = .324$. The group means were not statistically significantly different ($p = .324$). Therefore, the null hypothesis failed to be rejected.

Table 15

Welch ANOVA Fifth Grade 2018

	F	*df1*	*df2*	*Sig.*
Math2018	.348	2	70.534	.707

A one-way Welch ANOVA was conducted to determine the differences in the fifth grade proficient and advanced percentages of the 2018 math MAP scores among school districts using the three different evaluation systems (see Table 15). School districts were classified into three groups: MEES (n = 53), NEE (*n* = 175), and District (*n* = 36). There were no outliers as assessed by boxplot (see Appendix E); data was normally distributed for each group, as assessed by Shapiro-Wilk test ($p > .05$); but there was heterogeneity of variances, as assessed by Levene's test of homogeneity of variances ($p = .011$). The 2018 fifth grade math MAP proficient and advanced percentages increased from the MEES ($M = 37.5$, $SD = 17.8$), to District ($M = 38.8$, $SD = 17.4$), to NEE ($M = 39.7$, $SD = 13.8$), in the said order, but the differences between these evaluation systems was not statistically significant, Welch's $F(2, 70.534) = .348$, $p = .707$. The group means were not statistically significantly different ($p = .707$). Therefore, the null hypothesis failed to be rejected.

Research Question 7: What are the differences in the **third grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system and DESE approved, district developed evaluation model in **2019**?

$H_0$7. There are no statistically significant differences in the **third grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts

that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system and a DESE approved, district developed evaluation model in **2019**.

Table 16

ANOVA Third Grade 2019

		Sum of Squares	*df*	Mean Square	*F*	*Sig.*
ELA	Between Groups	194.488	2	97.244	.578	.561
	Within Groups	48428.299	288	168.154		
	Total	48622.786	290			
Math	Between Groups	739.048	2	369.524	1.774	.172
	Within Groups	55825.091	268	208.303		
	Total	56564.139.	270			

Table 17

Tukey Multiple Comparisons Third Grade 2019

Dependent Variable	(I) Eval	(J) Eval	Mean Difference (I-J)	Std. Error	*Sig.*	95% Confidence Interval Lower Bound	Upper Bound
ELA2019	MEES	NEE	.701	1.989	.934	-3.98	5.39
		District	-1.767	2.767	.799	-8.29	4.75
	NEE	MEES	-701	1.989	.934	-5.39	3.98
		District	-2.467	2.321	.538	-7.93	3.00
	District	MEES	1.767	2.767	.799	-4.75	8.29
		NEE	2.467	2.321	.538	-3.00	7.93
Math2019	MEES	NEE	-3.300	2.251	.309	-8.61	2.01
		District	-5.652	3.144	.172	-13.06	1.76
	NEE	MEES	3.300	2.251	.309	-2.01	8.61
		District	-2.352	2.663	.651	-8.63	3.92
	District	MEES	5.652	3.144	.172	-1.76	13.06
		NEE	2.352	2.663	.651	-3.92	8.63

A one-way ANOVA was conducted to determine the differences in the third grade proficient and advanced percentages of the 2019 ELA MAP scores among school districts that use the three different evaluation systems (see Table 16). School districts were classified into three groups: MEES (n = 54), NEE (n = 200), and District (n = 37). There were no outliers, as assessed by boxplot (see Appendix C); data was normally distributed for each group, as assessed by Shapiro-Wilk test (p > .05); and there was homogeneity of

variances, as assessed by Levene's test of homogeneity of variances (p = .130). The 2019 third grade ELA MAP proficient and advanced percentages increased from the NEE (M = 47.0, SD = 12.1), to MEES (M = 47.7, SD = 15.0), to District (M = 49.5, SD = 14.2), in the said order, but the differences between the evaluation systems was not statistically significant, $F(2, 288) = .578$, p = .561. The group means were not statistically significantly different (p = .561). Therefore, the null hypothesis failed to be rejected.

A one-way ANOVA was conducted to determine the differences in the third grade proficient and advanced percentages of the 2019 math MAP scores among school districts that use the three different evaluation systems (see Table 16). School districts were classified into three groups: MEES (n = 53), NEE (n = 183), and District (n = 35). There were no outliers, as assessed by boxplot (see Appendix F); data was normally distributed for each group, as assessed by Shapiro-Wilk test ($p > .05$); and there was homogeneity of variances, as assessed by Levene's test of homogeneity of variances (p = .692). The 2019 third grade math MAP proficient and advanced percentages increased from the MEES (M = 42.9, SD = 15.9), to NEE (M = 46.2, SD = 13.9), to District (M = 48.6, SD = 14.7), in the said order, but the differences between these evaluation systems was not statistically significant, $F(2, 268) = 1.774$, p = .172. The group means were not statistically significantly different (p = .172). Therefore, the null hypothesis failed to be rejected.

Research Question 8: What are the differences in the **fourth grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system and DESE approved, district developed evaluation model in **2019**?

$H_0$8. There are no statistically significant differences in the **fourth grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system and a DESE approved, district developed evaluation model in **2019**.

Table 18

Welch ANOVA Fourth Grade 2019

	F	*df1*	*df2*	*Sig.*
ELA2019	3.085	2	78.569	.051
Math2019	2.266	2	79.291	.110

Table 19

Tukey Multiple Comparisons Fourth Grade 2019

Dependent Variable	(I) Eval	(J) Eval	Mean Difference (I-J)	Std. Error	*Sig.*	95% Confidence Interval Lower Bound	Upper Bound
ELA2019	MEES	NEE	-3.982	2.057	.131	-8.83	.86
		District	-7.593*	2.912	.026	-14.45	-.73
	NEE	MEES	3.982	2.057	.131	-.86	8.83
		District	-3.611	2.485	.315	-9.46	2.24
	District	MEES	7.593*	2.912	.026	.73	14.45
		NEE	3.611	2.485	.315	-2.24	9.46
Math2019	MEES	NEE	-5.030	2.347	.083	-10.56	.50
		District	-4.173	3.250	.405	-11.83	3.49
	NEE	MEES	5.030	2.347	.083	-.50	10.56
		District	.857	2.743	.948	-5.61	7.32
	District	MEES	4.173	3.250	.405	-3.49	11.83
		NEE	-857	2.743	.948	-7.32	5.61

*. The mean difference is significant at the 0.05 level.

A one-way Welch ANOVA was conducted to determine the differences in the fourth grade proficient and advanced percentages of the 2019 ELA MAP scores among school districts using the three different evaluation systems (see Table 18). School districts were classified into three groups: MEES (n = 59), NEE (*n* = 200), and District (*n* = 37). There were no outliers as assessed by boxplot (see Appendix C); data was normally distributed for each group, as assessed by Shapiro-Wilk test ($p > .05$); but there

was heterogeneity of variances, as assessed by Levene's test of homogeneity of variances (p = .018). The 2019 fourth grade ELA MAP proficient and advanced percentages increased from the MEES (M = 43.4, SD = 16.9), to NEE (M = 47.4, SD = 13.1), to District (M = 51.0, SD = 12.8), in the said order, but the differences between these evaluation systems was not statistically significant, Welch's $F(2, 78.569) = 3.085$, p = .051. The group means were not statistically significantly different (p = .051). Therefore, the null hypothesis failed to be rejected.

Table 20

ANOVA Fourth Grade 2019

		Sum of Squares	*df*	Mean Square	*F*	*Sig.*
Math	Between Groups	1525.771	2	762.885	3.473	.033
	Within Groups	53602.787	244	219.684		
	Total	55128.558	246			

A one-way ANOVA was conducted to determine the differences in the fourth grade proficient and advanced percentages of the 2019 math MAP scores among school districts that use the three different evaluation systems (see Table 20). School districts were classified into three groups: MEES (n = 53), NEE (n = 183), and District (n = 36). There were no outliers, as assessed by boxplot (see Appendix F); data was normally distributed for each group, as assessed by Shapiro-Wilk test ($p > .05$); and there was homogeneity of variances, as assessed by Levene's test of homogeneity of variances (p = .533). The 2019 fourth grade math MAP proficient and advanced percentages increased from the MEES (M = 43.4, SD = 15.1), to District (M = 48.0, SD = 13.9), to NEE (M = 48.4, SD = 15.2), in the said order, but the differences between the evaluation systems was not statistically significant, $F(2, 244) = 3.473$, p = .033. The group means were not

statistically significantly different (p = .033). Therefore, the null hypothesis failed to be rejected.

Research Question 9: What are the differences in the **fifth grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system and DESE approved, district developed evaluation model in **2019**?

$H_0$9. There are no statistically significant differences in the **fifth grade** proficient and advanced percentages of the Missouri Assessment Program scores among school districts that use the Model Educator Evaluation System, the Network for Educator Effectiveness evaluation system and a DESE approved, district developed evaluation model in **2019**.

Table 21

ANOVA Fifth Grade 2019

		Sum of Squares	*df*	Mean Square	*F*	*Sig.*
ELA	Between Groups	517.899	2	258.949	1.572	.209
	Within Groups	47770.756	290	164.727		
	Total	48288.655	292			

Table 22

Tukey Multiple Comparisons Fifth Grade 2019

Dependent Variable	(I) Eval	(J) Eval	Mean Difference (I-J)	Std. Error	*Sig.*	95% Confidence Interval Lower Bound	Upper Bound
ELA2019	MEES	NEE	-2.965	1.928	.275	-7.51	1.58
		District	-4.297	2.710	.253	-10.68	2.09
	NEE	MEES	2.965	1.928	.275	-1.58	7.51
		District	-1.332	2.298	.831	-6.75	4.08
	District	MEES	4.297	2.710	.253	-2.09	10.68
		NEE	1.332	2.298	.831	-4.08	6.75
Math2019	MEES	NEE	-.193	2.399	.996	-5.85	5.46
		District	.950	3.341	.956	-6.92	8.82
	NEE	MEES	.193	2.399	.996	-5.46	5.85
		District	1.143	2.835	.914	-5.54	7.82
	District	MEES	-.950	3.341	.956	-8.82	6.92
		NEE	1.143	2.835	.914	-7.82	5.54

A one-way ANOVA was conducted to determine the differences in the fifth grade proficient and advanced percentages of the 2019 ELA MAP scores among school districts that use the three different evaluation systems (see Table 21). School districts were classified into three groups: MEES (n = 57), NEE (n = 199), and District (n = 37). There were no outliers, as assessed by boxplot (see Appendix C); data was normally distributed for each group, as assessed by Shapiro-Wilk test ($p > .05$); and there was homogeneity of variances, as assessed by Levene's test of homogeneity of variances ($p = .208$). The 2019 fifth grade ELA MAP proficient and advanced percentages increased from the MEES (M = 42.1, SD = 14.9), to NEE (M = 45.0, SD = 12.1), to District (M = 46.4, SD = 13.3), in the said order, but the differences between these evaluation systems was not statistically significant, $F(2, 290) = 1.572$, $p = .209$. The group means were not statistically significantly different ($p = .209$). Therefore, the null hypothesis failed to be rejected.

Table 23

Welch ANOVA Fifth Grade 2019

	F	*df1*	*df2*	*Sig.*
Math2019	.074	2	69.972	.929

A one-way Welch ANOVA was conducted to determine the differences in the fifth grade proficient and advanced percentages of the 2019 math MAP scores among school districts using the three different evaluation systems (see Table 23). School districts were classified into three groups: MEES (n = 53), NEE (n = 179), and District (n = 35). There were no outliers as assessed by boxplot (see Appendix F); data was normally distributed for each group, as assessed by Shapiro-Wilk test ($p > .05$); but there was heterogeneity of variances, as assessed by Levene's test of homogeneity of variances (p = .016). The 2019 fifth grade math MAP proficient and advanced percentages increased

from the District (M = 38.3, SD = 16.4), to MEES (M = 39.3, SD = 18.3), to NEE (M = 39.5, SD = 14.1), in the said order, but the differences between the evaluation systems was not statistically significant, Welch's $F(2, 69.972) = .074$, $p = .929$. The group means were not statistically significantly different ($p = .929$). Therefore, the null hypothesis failed to be rejected.

Summary

Chapter Four presented an analysis of the data. The research questions and null hypotheses were presented (see Table 24, 25, & 26). Samples were described and the results of the one-way ANOVA were shared. Each research question and null hypotheses was presented along with the data from the one-way ANOVA, Tukey's post hoc tests, and Welch ANOVAs when homogeneity of variances was violated. Each null hypothesis was rejected when the results yielded a statistically significant difference and effect size for the difference was calculated using Cohen's d. If the results did not yield a statistically significant difference, the null hypothesis could not be rejected.

Table 24

2017 Null Hypotheses Summary

	Null Hypothesis		Test	Sig.	Decision
$H_0$1	There are no statistically significant differences in the third grade proficient and advanced percentages of the MAP scores among school districts that use the MEES, the NEE, a district developed evaluation model in 2017.	ELA	Welch ANOVA	.109	Failed to reject the null hypothesis
		Math	Welch ANOVA	.149	Failed to reject the null hypothesis
$H_0$2	There are no statistically significant differences in the fourth grade proficient and advanced percentages of the MAP scores among school districts that use the MEES, the NEE, a district developed evaluation model in 2017.	ELA	ANOVA	.005	Reject the null hypothesis
		Math	ANOVA	.033	Reject the null hypothesis
$H_0$3	There are no statistically significant differences in the fifth grade proficient and advanced percentages of the MAP scores among school districts that use the MEES, the NEE, a district developed evaluation model in 2017.	ELA	ANOVA	.025	Reject the null hypothesis
		Math	ANOVA	.006	Reject the null hypothesis

Table 25

2018 Null Hypothesis Summary

	Null Hypothesis		Test	Sig.	Decision
$H_0$4	There are no statistically significant differences in the third grade proficient and advanced percentages of the MAP scores among school districts that use the MEES, the NEE, a district developed evaluation model in 2018.	ELA	ANOVA	.191	Failed to reject the null hypothesis
		Math	ANOVA	.487	Failed to reject the null hypothesis
$H_0$5	There are no statistically significant differences in the fourth grade proficient and advanced percentages of the MAP scores among school districts that use the MEES, the NEE, a district developed evaluation model in 2018.	ELA	ANOVA	.003	Reject the null hypothesis
		Math	ANOVA	.658	Failed to reject the null hypothesis
$H_0$6	There are no statistically significant differences in the fifth grade proficient and advanced percentages of the MAP scores among school districts that use the MEES, the NEE, a district developed evaluation model in 2018.	ELA	ANOVA	.324	Failed to reject the null hypothesis
		Math	Welch ANOVA	.707	Failed to reject the null hypothesis

Table 26

2019 Null Hypothesis Summary

	Null Hypothesis		Test	Sig.	Decision
$H_0$7	There are no statistically significant differences in the third grade proficient and advanced percentages of the MAP scores among school districts that use the MEES, the NEE, a district developed evaluation model in 2019.	ELA	ANOVA	.561	Failed to reject the null hypothesis
		Math	ANOVA	.172	Failed to reject the null hypothesis
$H_0$8	There are no statistically significant differences in the fourth grade proficient and advanced percentages of the MAP scores among school districts that use the MEES, the NEE, a district developed evaluation model in 2019.	ELA	Welch ANOVA	.051	Failed to reject the null hypothesis
		Math	ANOVA	.102	Failed to reject the null hypothesis
$H_0$9	There are no statistically significant differences in the fifth grade proficient and advanced percentages of the MAP scores among school districts that use the MEES, the NEE, a district developed evaluation model in 2019.	ELA	ANOVA	.209	Failed to reject the null hypothesis
		Math	Welch ANOVA	.929	Failed to reject the null hypothesis

In fourth grade ELA 2017 (see Table 24), the difference in proficient and advanced scores from the MEES to the NEE was statistically significantly different, indicating the NEE could have a greater impact on student achievement in ELA at the fourth grade level, but the effect size (d = .11) is small. In fourth grade math 2017, the difference in proficient and advanced scores from the MEES to the NEE were statistically significantly different, indicating the NEE could have a greater impact on student achievement in math at the fourth grade level, but the effect size (d = .12) is small. In fifth grade ELA 2017, the difference in proficient and advanced scores from the MEES to

the NEE were statistically significantly different, indicating the NEE could have a greater impact on student achievement in ELA at the fifth grade level, but the effect size (d = .10) is small. In fifth grade math 2017, the difference in proficient and advanced scores from the MEES to the NEE were statistically significantly different, indicating the NEE could have a greater impact on student achievement in math at the fifth grade level, but the effect size (d = .19) is small. In fourth grade ELA 2018 (see Table 25), the differences in proficient and advanced scores from the MEES to the NEE, as well as from the MEES to the District were statistically significantly different, indicating the NEE could have a greater impact on student achievement in ELA at the fourth grade level. The effect size for the difference from the MEES to the NEE (d = .14) is small, as is the effect size for the difference from the MEES to the District (d = .15).

Chapter Five will present the purpose of the quantitative, causal-comparative study and a summary of the methods used to complete the study. The researcher will share the findings of the research study, specifically any findings of statistically significant difference in the data from the results of the study. Based on the research findings, the researcher will share any conclusions and implications for practice, along with possible research topics with the intent those topics will extend the research involving evaluation systems and student achievement and add to the research in the future.

Chapter Five

Conclusions and Recommendations

Introduction

In his meta-analysis of the research into influences on student achievement, John Hattie (2009, 2012) identified the teacher as one of the most influential. Of the top twenty influences on student achievement, Hattie attributed four influences to the teacher and eight influences to the teaching. In consideration of Hattie's research, moving toward teacher evaluations that aid teachers in improving instruction is crucial to improving student achievement. The purpose of this quantitative, causal-comparative study was to determine the differences in the percent of students scoring proficient and advanced on the Missouri Assessment Program (MAP) in third, fourth, and fifth grades, depending on which evaluation system was used to evaluate the teacher between the years 2017 and 2019. The percent of students scoring proficient and advanced on the MAP for school districts in Missouri using the MEES, NEE, and a district created evaluation for the years 2017, 2018, and 2019 was retrieved from the MCDS portal at the MODESE open access website. A one-way ANOVA was conducted to compare the means to explore the differences in the proficient and advanced scores for each grade level among the evaluation systems.

Chapter Five presents a summary of the findings from Chapter Four followed by a discussion of the findings. Conclusions based on the findings of the research follow. The limitations of the study are also revisited. Implications for practice and recommendations for future studies are offered as well.

Summary of Findings

Educators agree there is a need for effective evaluation systems (Marzano, 2012; Danielson, 2011; Darling-Hammond, 2013), however agreement on what an effective evaluation system should include continues to be a source of debate (Grissom & Loeb, 2017). The research on how systems might impact student achievement is limited. To explore this gap in the research, the purpose of this study was to determine the difference in the percent of students scoring proficient and advanced on the MAP in third, fourth and, fifth grades, depending on which evaluation system was used to evaluate the teacher between the years 2017 and 2019.

Exploring the differences in the number of students scoring proficient and advanced on the MAP among school districts using the MEES, NEE, and a district created evaluation system, the study yielded six findings of statistical significance in three of the null hypotheses. In $H_0$2, a statistical difference was found in fourth grade ELA from the MEES to the NEE and in fourth grade math from the MEES to the NEE. In $H_0$3, a statistical difference was found in fifth grade ELA from the MEES to the NEE and in fifth grade math from the MEES to the district. Finally, in $H_0$5, a statistical difference was found in fourth grade ELA from the MEES to the NEE as well as from the MEES to the district. In all of the findings of statistical significance, the effect size was small.

Four of the six findings of statistical significance showed the difference in the percent of students scoring proficient and advanced to be from the MEES to the NEE. These occurred in 2017 fourth grade ELA and math MAP, 2017 fifth grade ELA MAP, and 2018 fourth grade ELA MAP. While this could indicate the NEE has a greater impact on student achievement than the MEES, the effect size in each of the comparisons was

low ($d < .12$). Despite the finding of statistical significance, these findings are of little practical use due to the small effect size. There are, however, some implications which can be considered when the data is explored in relation to what is known about teacher evaluation.

Considering MODESE's (2013c) efforts to create the MEES, a research-based evaluation system developed by educators and based on the work of Robert Marzano, John Hattie, Charlotte Danielson, and Doug Lemov, it would have been expected students whose teachers were evaluated using the MEES would have performed better. In the study, a total of fifty-four comparisons were made from the nine null hypotheses. Of those, only six showed statistical significance. In all six, the MEES was shown to have the lowest percent of students scoring proficient and advanced. When applying the Rationalistic Theory (Bhola, 1990; Darling-Hammond & Wise, 1981), which has been applied in the educational setting to predict outcomes on assessments, the pattern would lead to the prediction that teachers evaluated with the MEES will continue to have students who do not perform as well on the MAP as students whose teachers are evaluated with the NEE. This prediction is confirmed when considering the remaining two findings of statistical significance as well as the findings that were not of statistical significance, yet still provided useful information to the research.

Two of the six findings of statistical significance demonstrated the difference in the percent of students scoring proficient and advanced to be from teachers evaluated with the MEES to teachers evaluated with a district created evaluation system. The two findings occurred in the 2017 fifth grade math MAP and the 2018 fourth grade ELA MAP. While this could indicate a district created evaluation system has a greater impact

on student achievement than the MEES, the effect size in each of the comparisons was low (d = < .19). Despite the findings of statistical significance, these findings are of little practical use due to the small effect size. There are, however, some implications that can be considered when the data is explored in relation to what is known about teacher evaluation.

As with the four findings of statistical significance from the MEES to the NEE, it would have been expected students whose teachers were evaluated using the MEES would have performed better. One explanation for the statistical significance from the MEES to the NEE could be supported by the research of Feingold (2013) and Jennings and Sohn (2014) regarding development of evaluations systems. Feingold (2013) and Jennings and Sohn (2014) suggest including teachers in the creation of an evaluation system, results in more effective evaluations. Including teachers in the creation of an evaluation system and the result of a more effective evaluation is likely due to the fact teachers are closest to the work, interacting with the students, curriculum, and the Missouri Learning Standards, on which the MAP is based. Again, supporting the Rationalistic Theory as it can be predicted teachers evaluated with the MEES will continue to have students who do not perform as well as students whose teachers are evaluated using the NEE or a district created evaluation.

The study had nine null hypotheses, each comparing ELA MAP scores and math MAP scores, for three evaluation systems. The results yielded fifty-four total comparisons of students scoring proficient and advanced among schools using the MEES, NEE, and a district created evaluation system. While only six of those comparisons showed statistical significance in the difference of students scoring proficient and

advanced, an exploration of the data including findings, which were not of statistical significance, still provided valuable information. The nine null hypotheses presented the mean percent of students scoring proficient and advanced in ELA and math for each grade, each year. Of the nine hypotheses, only two, 2019 third grade ELA and 2019 fifth grade math, indicated the MEES did not have the lowest mean percent of students scoring proficient and advanced. While having the lowest percent of students scoring proficient and advanced does not indicate the MEES is the least effective evaluation system, this information, combined with the fact that all six findings of statistical significance were from the MEES to either the NEE or a district created evaluation system could be of value to leaders when researching and making decisions about which evaluation system to implement. With three years of data showing no significant growth among students whose teacher is evaluated using the MEES, MODESE does not appear to be meeting the goal of providing an effective teacher evaluation to provide the best education for students and meet the guidelines for students as outlined in NCLB and ESSA (MODESE, 2015).

Conclusions

The purpose of this quantitative, causal-comparative study was to determine the differences in the percent of students scoring proficient and advanced on the MAP in third, fourth, and fifth grades, depending on which evaluation model was used to evaluate the teacher between the years 2017 and 2019. While there were only six findings of statistical significance, patterns emerged in the data that, when analyzed, yielded unexpected results. It also became apparent the Rationalistic Theory of student achievement applied in the study and Marzano's theory of teacher evaluation applied in

the study appeared to be contradictory, presenting a challenge.

Marzano's theory of evaluation, the Focused Teacher Evaluations, states "the purpose of supervision should be the enhancement of teachers' pedagogical skills, with the ultimate goal of enhancing student achievement" (2011, p. 2). In the study, out of fifty-four comparisons, there were only six findings of statistical significance, all with an effect size small enough to render them of little practical use on their own. However, when considered with the findings not of statistical significance, some important information could be found leading to conclusions, which are useful for educational implications in both evaluation and student achievement. An unexpected result of the data was the performance of students whose teachers were evaluated using the MEES, the MODESE created evaluation. Of the fifty-four comparisons, twenty-seven comparisons for ELA and twenty-seven comparisons for math over the three years, in three grade levels, the MEES had the lowest mean percent of students scoring proficient and advanced; twenty-six times in ELA and twenty-six times in math. Considering the work MODESE put into creating the MEES, the research model was based on, including Marzano's strategies and Hattie's meta analyses of student achievement, it was expected the MEES would have yielded higher results.

Another unexpected result of the study was the performance of district created evaluation systems. District created systems were developed independently by individual school districts, which would most likely indicate distinct differences when compared to the evaluation systems of other districts across the state. Despite each district created evaluation system's unique characteristics, district created systems had the highest mean percent of students scoring proficient and advanced. Of the fifty-four comparisons,

district created systems had the highest mean percent ten times; five in ELA and five in math. While these results do not indicate a district created system is the best evaluation system to use, when considered along with the findings of significance, district created evaluations can indicate the need for further exploration into district created systems, and how such evaluations might impact student achievement. Although district created evaluation systems are developed by teachers and administrators of each individual school district, each school district may not have access to the same resources to help aid in the development of an evaluation system. Research has suggested the inclusion of educators in the creation of evaluations, results in more a more effective evaluation process (Feingold, 2013; Jennings & Sohn, 2014). Districts who created their own evaluations benefited from the experience and knowledge of their staff, in addition to the research MODESE provided as a guideline for districts to follow in the creation of the evaluation. The results of the study provide additional understanding regarding the development of evaluations for leaders at both the district and state levels. for educational leaders at the district and state level in the development of evaluations.

Overall, the results of the study revealed teacher evaluation alone is likely to have very little impact on student achievement due to two factors: the complex nature of teaching (Danielson, 2007; Marzano, 2007) and the continued practice of evaluators to rate teachers with high ratings (Sawchuk, 2013; Kraft & Gilmour, 2017). Hattie's meta-analysis in 2009 revealed 12 of the top twenty influences on a student's academic success are attributed to the teacher or the teaching. Hattie's work would indicate the art of teaching would be far too complex to capture in an evaluation, and even more difficult to be comprehensive enough to encourage change in a teacher's practice. Grissom and Loeb

(2017) asserted due to so many components in teaching, an evaluation cannot possibly assess each component. Grissom and Loeb's (2017) research echoes the work of researchers (Hattie, 2009; McDonnell, 2013; Galey, 2015) who suggest there are many variables which impact student achievement, which cannot be evaluated, and it is impossible to attribute a student's achievement solely based on an evaluation of the teacher. Some of those variables; socio-economic status, homelessness, attendance, family involvement, and student motivation, may not be able to be overcome, even by a quality teacher. Baker, et al. (2013), suggest holding a teacher accountable for student learning in situations, which cannot be controlled, is unreasonable and could be a violation of the rights of the teacher.

Another conclusion reached from the results of the study not directly tied to the data, was the contradiction in the theoretical frameworks. Because there was not one single theory, which pertained to both student achievement and teacher evaluation, two theories were applied in this study. According to Bhola (1990) and Darling-Hammond and Wise (1981), the Rationalistic Theory as applied to student achievement, is a logical theory that describes the student as a receiver of what the teacher delivers. Darling-Hammond and Wise (1981) described the classroom setting as one where predictions can be made based on past performance and those results can then be repeated with future students if the teacher delivers the instruction the same way; the oversimplification is all students learn in the same way. The theoretical framework applied to teacher evaluation was based on Marzano's theory of teacher evaluation (Marzano, 2012). Marzano's evaluation based on feedback provided to the teacher, related not just to teacher performance but also to evidence of the student achievement. Marzano (2012) suggests,

when students do not show evidence of achievement, feedback given to the teacher can help the teacher make changes so the student can achieve. Feedback can include providing professional coaching for the teacher or recommending professional development (Marzano, 2012). The conclusion presented a challenge but added to the research indicating the need for more research and theoretical frameworks, which pertain to both student achievement and teacher evaluation. The Gates Foundation has started some work with teacher evaluations and student achievement with the Measures of Effective Teaching (2010), however additional research is needed.

While there is increasing pressure on states to improve student achievement and provide more rigorous evaluation of teachers, there is a lack of research surrounding how teacher evaluations impact student achievement. This study added to the gap in the literature in three ways. First, the data revealed the MODESE created the MEES to be the evaluation system, which would be least likely to impact student achievement. A district created evaluation, based on the same standards as the MEES, was the evaluation system most likely to have an impact on student achievement. The conclusions drawn indicated evaluations should include the educators interacting with the students, the curriculum, and the standards on a daily basis. Second, the study revealed teacher evaluation alone are unlikely to impact student achievement. Finally, the study added to the gap in the literature by identifying the need for theories that apply to student achievement and teacher evaluation and their link to NCLB and ESSA.

Limitations Indicated

Limitations of the research study may have impacted the results of the study. The research was limited to students in grades three, four, and five in public schools in the

state of Missouri with private or parochial schools excluded. The results of this study would not apply to students outside of the defined parameters, which included the evaluations used. Another limitation of the study is the study only addressed how the use of three Missouri approved evaluation systems may impact student achievement and did not consider many other variables, which may impact student achievement. The results of the study can only determine a relationship between the evaluation tool used and student achievement, not causation. There are limitations in the validity and reliability of the MAP test; however, Missouri has made improvements in the area of reliability with the change to the DRC for the MAP assessment. Lastly, a limitation, which cannot be controlled, but could impact the results of the study, was how strictly the school district personnel adhere to the guidelines and requirements of the evaluation system.

Implications for Practice

Marzano's (2013) framework for evaluation shifted the focus of teacher evaluations comprised of scripted observations to focused more on identifying strategies the teacher can implement to positively impact student achievement. Using feedback from teachers on specific strategies and how those strategies are applied, evaluators can better assist teachers in improving instruction and increase student achievement. The use of research-based strategies, as recommended by Marzano (2013), can potentially impact teacher evaluation results.

Teacher evaluations used as an effective tool can help the teacher grow and develop as an effective educator and impact student achievement. The education community is obligated to put real work into ensuring teacher evaluations are a useful tool and not simply an act of compliance with a political mandates. District created

evaluations should ensure both the evaluator and the teacher find the process of creating and then implementing the evaluation beneficial. Including the teacher in the creation of the evaluation should be a priority since use of the evaluation can potentially impact a teacher's instruction and student achievement. Teachers are more likely to be invested if included in the evaluation process and may likely accept feedback from the evaluator as a result. Researchers also suggest creating evaluations specific to the discipline or subject area being taught. The implication of specific evaluations could help to provide teachers with feedback, which could be more focused on specific needs, and promote more willingness from the teacher to implement recommended changes. Another suggestion when creating evaluations is the consideration of special circumstances such as classrooms with a high poverty rate, homeless populations or transient rate, all of which could have a significant impact on student achievement, decreasing the impact of the teacher.

Authenticity in the evaluation process is an area that should be improved. One area in which authenticity needs improvement is making evaluations more purposeful instead of performing evaluations for the sake of to be compliance. Examples of how to make evaluations more authentic include developing an evaluation tool that is applicable to the teacher, and including meaningful and purposeful ratings for the teacher as well. Despite reforms and new evaluations, many teachers continue to receive ratings in the top categories regardless of skill level. Receiving top ratings does not serve to improve student achievement, as poor quality teachers may continue to teach with a false sense of accomplishment while high quality teachers may not receive recognition for standing out as exemplary. Finally, evaluators need to prioritize teacher evaluations for the purpose of

improving teacher practice and impacting student achievement and not complete evaluations for the purpose of fulfilling a requirement.

Recommendations for Future Studies

Considering the results of this study, it is evident more information is needed to determine how the teacher evaluation process impacts student achievement. Only six of the fifty-four comparisons made in the nine null hypotheses yielding a statistically significant difference in the achievement of students among the three evaluation systems used. None of the results of statistical significance resulted in more than a small effect size. The study did not consider the demographics of the schools included in the study so future researchers building on this study may find it beneficial to consider the data based on a demographic representation. Additionally, researchers could look at the results based on the differences in teacher tenure. Another possibility for future studies could include a demographic study of the schools based on the differences in location; urban, rural, and suburban school districts. Finally, third possibility of additional research could focus on a demographic study based on school size, comparing the results based on student population.

Future research could look more closely at how teacher evaluation impacts change of a teachers' pedagogy. A perception study based on teachers' perceptions of how current teaching practices may change based on the evaluation system used by the evaluator. Such a study could yield insight into how pedagogical practice changes as well as why students achieve at high levels. Additionally, a principals' perception study based on their perceptions of how their teachers practice changes based on the evaluation process used could also yield insight into why student achievement may or may not

increase.

A final thought on future research studies relates specifically to the Missouri Educator Evaluation Model. While data was not collected on the number of schools that changed evaluation models, it was noted by the researcher there were schools that changed from using the MEES to either the NEE or a districted created model. Additionally, the results of the study did not reveal the MEES to have a greater impact on student achievement than either the NEE or a district created model. It was observed that the MEES had the lowest percent of students scoring proficient and advanced when compared with the NEE and district created evaluations in sixteen of the eighteen comparisons. Understanding why the evaluation system has not shown a stronger effect on student achievement would be beneficial in future research. Continued research could yield insight into successful development of evaluations that impact students and could include a perceptual study of schools that switched evaluation systems, specifically indicating the reasons the initial evaluation system was inadequate.

Summary

States continue to seek ways to improve student achievement. Comparisons continue to be made across the United States and among students in other countries. Unfortunately, the publication of "A Nation at Risk" in 1983 continues to be a topic of conversation. States have responded to federal mandates and put in place new assessments to measure student achievement and developed new evaluations to measure teacher quality. According to Marzano, the purpose of the teacher evaluation was to improve a teacher's skills to improve student achievement (2011). With the focus of improving teacher evaluation and student achievement in mind, the purpose of this

causal-comparative study was to explore the differences in the percent of students scoring proficient and advanced on the MAP in third, fourth and fifth grades, depending on which evaluation model was used to evaluate the teacher between the years 2017 and 2019.

This study explored the differences in three teacher evaluation systems used in Missouri. The evaluation systems were the MEES, NEE, and a district created evaluation system. Statistically significant results were found in three of the nine null hypotheses. These results were in H_o2, for fourth grade ELA and fourth grade math in 2017, H_o3 for fifth grade ELA and fifth grade math in 2017, and H_o5 for fourth grade ELA in 2018. In each of the tests of statistical significance, the effect size was small ($d < .20$). While the results are of statistical significance, the small effect size renders the results of minimal practical use. Though the results had a small effect size, a clear pattern did emerge. In each of the findings of statistical significance, the pattern that emerged was the increase in the mean percent of students scoring proficient and advanced from teachers evaluated using the MEES to teachers evaluated using the NEE. This statistical significance occurred for the 2017 4th grade ELA, 2017 4th grade math, 2017 5th grade ELA, 2017 5th grade math, and 2018 4th grade ELA. In 2018 ELA, the difference from the MEES to the district created evaluation system also showed statistical significance. Results that occurred in 2017 were one year before the MAP changed and a new test was implemented. Because of the transition to a new test, MODESE guards against comparing the current MAP results to the results of MAP assessments administered after 2017. Overall, the results of this quantitative, causal comparative study showed there to be very little difference in the MEES, NEE, and a district created evaluation system, the three evaluation systems used by most schools in Missouri. Even in the six results of

statistical significance, the effect size was small, rendering if of minimal practical use.

The researcher found several patterns in the overall data overall data of the study. Patterns in the data yielded unexpected results, which could provide insight for the future development of teacher evaluations as well as future research. Teachers evaluated with the MODESE developed MEES had the lowest mean percent of students scoring proficient and advanced on the MAP in sixteen of the fifty-four comparisons, while teachers evaluated using a district created evaluation, had the highest mean percent of students scoring proficient and advanced on ten of the fifty-four comparisons. While the results do not indicate the MEES is the least effective evaluation system, the information, is of value to leaders when making decisions about evaluation systems.

The theoretical frameworks applied in the study were contradictory in regard to the application of each framework in the study. The Rationalistic Theory (Darling-Hammond & Wise, 1981; Bhola, 1990) applied to student achievement stated behavior could be affected if a specific set of actions is applied and predictions can be made about how students with similar characteristics will perform when placed in similar situations. The MAP overgeneralizes the Rationalistic Theory with the expectation that all students in the same grade level will perform the same on the MAP, which is given at the same time every year. Marzano's (2012) theory emphasizing the purpose of evaluation of teachers should enhance the teacher's skills in an effort to improve the students' achievement is inconclusive from the results of the study. Each of the findings of statistical significance had a small effect size, which makes the results of the findings to be of minimal practical use. Additional research could be conducted to determine a correlation between the evaluation and student achievement, however causation would be

difficult, if not impossible, to isolate.

This study contributed to the research by exploring the differences in the achievement of students in Missouri based on the evaluation system used to evaluate the teachers over the last three years. The intent of the research conducted was to determine any statistically significant differences in the achievement of the students. The information could inform leaders when making decisions regarding which evaluation to use in an effort to influence student achievement.

The need for teacher evaluation and measures of student achievement are not in dispute. As with any profession, it is necessary to provide feedback to employees on whether job performance meets expectations, exceeds expectations, or falls short. Unlike most professions, the outcome of a teacher's work, student achievement, cannot be easily measured and is impacted by variables outside of the teacher's control. According to the Rationalistic Theory, similar students should respond in the same way to the same strategy and outcomes of achievement can be predicted. However, similarities cannot be controlled; teachers may not have knowledge of the specific variables which impact student achievement. The Rationalistic Theory would indicate the teacher should continue with the same strategy, however the Marzano's theory of evaluation would indicate the teacher should adjust for the student if the strategy is not working. The challenge in education exists in the difficulty of effectively evaluating teachers solely based on the variables over which they have control, and determining the impact specific variables have on the achievement of students.

References

Baker, B. D., Oluwole, J. O., & Green, P. C., III. (2013). The legal consequences of mandating high stakes decisions based on low quality information: Teacher evaluation in the race-to-the-top era. *Education Policy Analysis Archives*, *21*(5). Retrieved from http://search.ebscohost.com/login.aspx?direct=true&AuthType=shib&db=eric&AN=EJ1015318&site=eds-live

Bates, A., Shifflet, R., & Lin, M. (2013). Academic achievement: An elementary school perspective. In J. Hattie & E. M. Anderman, *International guide to student achievement* (pp. 7-9). New York, NY: Routledge.

Benedict, A. E., Thomas, R. A., Kimerling, J., & Leko, C. (2013). Trends in teacher evaluation. *Teaching Exceptional Children*, *45*(5), 60–68. https://doi.org/10.1177/004005991304500507

Bhola, H. S. (1990). *A Model of evaluation planning, Implementation and management toward a "culture of information" within organizations*. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&AuthType=shib&db=eric&AN=ED328590&site=eds-live

Blumberg, A. (1985). Where we came from: Notes on supervision in the 1840s. *Journal of Curriculum & Supervision*, *1*(1), 56–65. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&AuthType=shib&db=eue&AN=48151933&site=eds-live

Brevetti, M. (2014). Reevaluating narrow accountability in American schools: The need

for collaborative effort in improving teaching performances. *Delta Kappa Gamma Bulletin, 81*(1), 32-35. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&AuthType=shib&db=aph&AN=98474484&site=eds-live

Bruce, P. C. (2015). *Introductory statistics and analytics: A resampling perspective*. Hoboken, NJ: Wiley.

Burke, P. J., & Krey, R. D. (2005). *Supervision: A guide to instructional leadership*. Springfield, IL: Charles C. Thomas.

Carbaugh, B., Marzano, R., & Toth, M. (2017). *The Marzano focused teacher evaluation model: A focused, scientific-behavioral evaluation model for standards-based classrooms.* Learning Sciences International.

Cubberley, E. P. (1929) *Public school administration.* Retrieved from https://archive.org/details/publicschooladmi1922cubb/

Danielson, C. (2007). *Enhancing professional practice: a framework for teaching*. Alexandria, VA: Association for Supervision and Curriculum Development.

Danielson, C. (2011). Evaluations that help teachers learn. *The Effective Educator*, *68*(4), 35–39. Retrieved from http://www.ascd.org/publications/educational-leadership/dec10/vol68/num04/Evaluations-That-Help-Teachers-Learn.aspx

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

Darling-Hammond, L., Gendler, T., & Wise, A. E. (1990). *The teaching*

internship. Practical preparation for a licensed profession. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&AuthType=shib&db=eric&AN=ED326512&site=eds-live

Darling-Hammond, L. & Wise, A. E. (1981). *A conceptual framework for examining teachers' views of teaching and educational policies.* Santa Monica, CA: Rand Corporation.

Data Recognition Corporation. (2018). *Missouri Assessment Program grade-level assessments technical report.* Maple Grove, MN: Retrieved from file:///Users/teresaadams/Desktop/DEC29/asmt-gl-tech-report-2018.pdf

Dee, T., & Wyckoff, J. (2017). A lasting impact: High-stakes teacher evaluations drive student success in Washington, D.C. *Education Next, 17*(4), 58-66. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&AuthType=shib&db=eue&AN=125076084&site=eds-live

Donaldson, M. L. (May 2016). Teacher evaluation reform focus, feedback, and fear. *Educational Leadership. (73)*8, 72-77.

Donaldson, M. L. (2013). Principals' approaches to cultivating teacher effectiveness: Constraints and opportunities in hiring, assigning, evaluating, and developing teachers. *Educational Administration Quarterly, 49*(5), 838-882. doi: 10.1177/0013161X13485961

Emory, R., Caughy, M., Harris, R., & Franzini, L. (2008). Neighborhood social processes and academic achievement in elementary school. *Journal of Community Psychology, 36,* 886-896.

Ensuring Fair and Reliable Measures of Effective Teaching: Culminating Findings from the MET Project's Three-Year Study (2013). Policy and Practice Brief. MET Project. *Bill & Melinda Gates Foundation.* Retrieved September 25, 2019, from http://www.metproject.org/downloads/MET_Ensuring_Fair_and_Reliable_Measures_Practitioner_Brief.pdf

Every Student Succeeds Act: Federal Elementary and Secondary Education Policy. (2017). *Congressional Digest, 96*(7), 4-6. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&AuthType=shib&db=aph&AN=124783297&site=eds-live

Feingold, R. (2013). Vision in an age of accountability. *Quest (00336297), 64*(4), 385-393. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&AuthType=shib&db=s3h&AN=91536937&site=eds-live

Galey, S. (2015). Education politics and policy: Emerging institutions, interests, and ideas. *The Policy Studies Journal*, 43(1), 12-39.

Gay, L. R., Mills, G. E., Airasian, P. (2009). *Educational research: Competencies for analysis and applications* (9th ed.). Upper Saddle River, NJ: Pearson Education, Inc.

Graziano, S. K. (2017). *An exploration of teacher perception of the marzano causal teacher evaluation model and its impact on professional practices* (Order No. 10254663). Available from ProQuest Dissertations & Theses Global. (1870036820). Retrieved from https://search.proquest.com/docview/1870036820?accountid=14196

Grissom, J. & Loeb, S. (2017). Assessing principals' assessments: Subjective evaluations of teacher effectiveness in low- and high-stakes environments. *Education Finance and Policy, 12*(3), 369-395.

Guskey, T. R. (2013). Defining student achievement. In J. Hattie & E. M. Anderman, *International guide to student achievement* (pp. 3-6). New York, NY: Routledge.

Hanushek, E. A. (2016). What matters for student achievement: Updating Coleman on the influence of families and schools. *Education Next, 16*(2), 18-26. Retrieved from http://eds.a.ebscohost.com/eds/pdfviewer/pdfviewer?vid=3&sid=e0cd64da-28ec-4863-a30d-7d7f2fecb3f5%40sdc-v-sessmgr01

Hattie, John. (2009). *Visible learning*. Abingdon, Oxon: Routledge.

Hattie, John. (2012). *Visible learning for teachers: Maximizing impact on learning.* London; New York: Routledge.

Holland, P. E., & Garman, N. B. (2001). Toward a resolution of the crisis of legitimacy in the field of supervision. *Journal of Curriculum & Supervision*, *16*(2), 95–111. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&AuthType=shib&db=eue&AN=507730082&site=eds-live

Jacob, A. (2012). Examining the relationship between student achievement and observable teacher characteristics: Implications for school leaders. *International Journal of Educational Leadership Preparation*, *7*(3). Retrieved from http://search.ebscohost.com/login.aspx?direct=true&AuthType=shib&db=eric&AN=EJ997469&site=eds-live

Jennings, J., & Sohn,. H. (2014). Measure for measure: How proficiency-based

accountability systems affect inequality in academic achievement. *Sociology of Education. 87*(2), 125-141. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&AuthType=shib&db=eric&AN=EJ1021008&site=eds-live

Jewell, J. W. (2017). From inspection, supervision, and observation to value-added evaluation: A brief history of U.S. teacher performance evaluations. *Drake Law Review*, *65*, 363. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&AuthType=shib&db=edslex&AN=edslexFE25BE7D&site=eds-live

Kirkpatrick, M. B. (2010). Principals' perceptions on frustration, obstacles and change (Order No. 3411304). Available from Education Database. (519054964). Retrieved from https://search.proquest.com/pqdtglobal/docview/519054964/fulltextPDF/13AF5DF981E14063PQ/1?accountid=14196

Koyama, J., & Kania, B. (2014). When transparency obscures: The political spectacle of accountability. *Journal for Critical Education Policy Studies, 12*(1), 143-169. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&AuthType=shib&db=eric&AN=EJ1033454&site=eds-live

Kraft, M. A., & Gilmour, A. F. (2016). Can principals promote teacher development as evaluators? A case study of principals' views and experiences. *Educational Administration Quarterly*, *52*(5), 711–753. https://doi.org/10.1177/0013161X16653445

Laerd Statistics (2016). One-way ANOVA using strata. *Statistical tutorials and software guides*. Retrieved from https://statistics.laerd.com/

LaRocque, M., Kleiman, I., & Darling, S. M. (2011). Parental involvement: The missing link in school achievement. *Preventing School Failure: Alternative Education for Children and Youth, 55*(3), 115-122.

Lewis, G. W. (2018). *Elementary principal and teacher perceptions of the quality and accuracy of teacher evaluation ratings: A causal-comparative study* (Order No. 10811566). Available from ProQuest Dissertations & Theses Global. (2088102042). Retrieved from https://search.proquest.com/docview/2088102042?accountid=14196

Long, A. E. (2019). *Teachers' perceptions and experiences with a reformed teacher evaluation system: Conditions necessary for changing practice* (Order No. 13917937). Available from ProQuest Dissertations & Theses Global. (2246451544). Retrieved from https://search.proquest.com/docview/2246451544?accountid=14196

Mantzicopoulos, P., Patrick, H., Strati, A., & Watson, J. S. (2018). Predicting kindergarteners' achievement and motivation from observational measures of teaching effectiveness. *Journal of Experimental Education*, *86*(2), 214–232. https://doi.org/10.1080/00220973.2016.1277338

Mausethagen, S. (2013). Accountable for what and to whom? Changing representations and new legitimation discourses among teachers under increased external control. *Journal of Educational Change, 14*(4), 423-444. Retrieved from

http://search.ebscohost.com/login.aspx?direct=true&AuthType=shib&db=eric&AN=EJ1038190&site=eds-live

Marzano, R., (2007). *The art and science of teaching: A comprehensive framework for effective instruction.* Alexandria, VA: The Association for Supervision and Curriculum Development (ASCD).

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

Marzano, R., Frontier, T., & Livingston, D. (2011). *Effective supervision: Supporting the art and science of teaching.* Retrieved from http://www.ascd.org/publications/books/110019/chapters/A-Brief-History-of-Supervision-and-Evaluation.aspx

Maslow, V. J., & Kelley, C. J. (2012). Does evaluation advance teaching practice? The effects of performance evaluation on teaching quality and system change in large diverse high schools. *Journal of School Leadership, 22*(3), 600-632). Retrieved from https://eric.ed.gov/?id=EJ986805

McDonnell, L.M. (2013). Educational accountability and policy feedback. *Educational Policy, 27*(2), 170-189. doi: 10.1177/0895904812465119

McLeod, S. (2019). What does effect size tell you? Retrieved from https://www.simplypsychology.org/effect-size.html

Measures of Effective Teaching (MET) Project. (2010c, December). Learning about teaching: Initial findings from the Measures of Effective Teaching Project. Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from http://www.metproject.org/downloads/Preliminary_FindingsResearch_Paper.p

Mette, I. M., Range, B. G., Anderson, J., Hvidston, D. J., Nieuwenhuizen, L., & Doty, J. (2017). The wicked problem of the intersection between supervision and evaluation. *International Electronic Journal of Elementary Education, (3)*709. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&AuthType=shib&db=edsdoj&AN=edsdoj.99a5ba7bbdb943e682de337ef7af4615&site=eds-live

Missouri Department of Elementary and Secondary Education. (2012a). *Research and proven practices of Dr. John Hattie* [PDF file]. Retrieved from https://dese.mo.gov/sites/default/files/10-Research-ProvenPracticesHattie.pdf

Missouri Department of Elementary and Secondary Education. (2012b). *Research and proven practices of Dr. Robert Marzano* [PDF file]. Retrieved from https://dese.mo.gov/sites/default/files/09-Research-ProvenPracticesMarzano.pdf

Missouri Department of Elementary and Secondary Education. (2013a). *Essential principles of effective evaluation Missouri's educator evaluation system* [PDF file]. Retrieved from http://www.dese.mo.gov/eq/ees.htm

Missouri Department of Elementary and Secondary Education. (2013b). *Executive summary Missouri's educator evaluation system* [PDF file]. Retrieved from http://www.dese.mo.gov/eq/ees.htm

Missouri Department of Elementary and Secondary Education. (2013c). *Teacher evaluation Missouri's educator evaluation system* [PDF file]. Retrieved from http://www.dese.mo.gov/eq/ees.htm

Missouri Department of Elementary and Secondary Education. (September, 2014). *Missouri assessment program updates and changes.* [PDF file]. Retrieved from

https://dese.mo.gov/sites/default/files/asmt-map-report-to-legislature-1415.pdf

Missouri Department of Elementary and Secondary Education. (2015). *Missouri's ESEA flexibility waiver* [Brochure]. Retrieved from https://dese.mo.gov/sites/default/files/qs-esea-waiver-brochure.pdf

Missouri Releases Statewide Assessment Results. (2019). Retrieved from https://dese.mo.gov/communications/news-releases/missouri-releases-statewide-assessment-results-0

Mo. Rev. Stat. §168.128 (2014).

Network for Educator Effectiveness - The largest, most comprehensive teacher evaluation system in Missouri. (2017, January 11). Retrieved from https://education.missouri.edu/2015/11/network-for-educator-effectiveness-the-largest-most-comprehensive-teacher-evaluation-system-in-missouri/.

No Child Left Behind Act of 2002. Pub.L. No. 100-1000, (2002).

Papay, J. P. (2012). Refocusing the debate: Assessing the purposes and tools of teacher evaluation. *Harvard Educational Review*, *82*(1), 123–141. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&AuthType=shib&db=eric&AN=EJ969494&site=eds-live

Parker, N. B. (2017). *Framework for teacher evaluation: Examining the relationship between teacher performance and student achievement* (Order No. 10619652). Available from ProQuest Dissertations & Theses Global. (1952703543). Retrieved from https://search.proquest.com/docview/1952703543?accountid=14196

Pil, F., & Leana, C. (2009). Applying organizational research to public school reform:

The effects of teacher human and social capital on student performance. *Academy of Management Journal, 52,* 1101-1124.

Piro, J. S., & Mullen, L. (2013). Output as educator effectiveness in the United States: Shifting towards political accountability. *International Journal of Educational Leadership Preparation, 8*(2), 59-77. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&AuthType=shib&db=eric&AN=EJ1016271&site=eds-live

Range, B. G. (2013). How teachers perceive principal supervision and evaluation in eight elementary schools. *Journal of Research in Education, 23*(2), 65-78. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&AuthType=shib&db=eric&AN=EJ1098433&site=eds-live

Ritter, G. W. & Shuls, J. V. (2012). If a tree falls in a forest, but no one hears . . . *The Phi Delta Kappan*, *94*(3), 34. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&AuthType=shib&db=edsjsr&AN=edsjsr.41763673&site=eds-live

Sadeghi, L., & Callahan, K. (2013). Shifting educational accountability from compliance to outcomes. *Public Manager, 42*(3), 62-64. Retrieved from https://search.proquest.com/docview/1448425394?accountid=14196

Sandilos, L. E., Sims, W. A., Norwalk, K. E., & Reddy, L. A. (2019). Converging on quality: Examining multiple measures of teaching effectiveness. *Journal of School Psychology*, *74*, 10–28. https://doi.org/10.1016/j.jsp.2019.05.004

Sawchuk, S. (2013). Teachers' ratings still high despite new measures. *Education Week,*

32(20), 18-19. Retrieved from https://www.edweek.org/ew/articles/2013/02/06/20evaluate_ep.h32.html

Senechal, D. (2013). Measure against measure: Responsibility versus accountability in education. *Arts Education Policy Review. 114*(2), 47-53. https://doi.org/10.1080/10632913.2013.769828

Simon, M. & Goes, J. (2013). Ex post facto research: Using existing data for your dissertation research. Retrieved from http://www.dissertationrecipes.com/page/7/

Sporte, S. E., Jiang, J. Y., Luppescu, S., & Society for Research on Educational Effectiveness (SREE). (2016). Teacher evaluation in practice: Exploring relationships between school characteristics & evaluation scores. *Society for Research on Educational Effectiveness*. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&AuthType=shib&db=eric&AN=ED567738&site=eds-live

Strong, M., Gargani, J., & Hacifazlioglu, O. (2011). Do we know a successful teacher when we see one? Experiments in the identification of effective teachers. *Journal of Teacher Education*, *6*(4), 367-382.

Superfine, B.M., Gottlieb, J.J., & Smylie, M.A. (2012). The expanding federal role in teacher workforce teacher policy. *Educational Policy, 27*(1), 58-78. doi: 10.1257/ger.102.7.3628

Taylor, E.S., & Tyler, J.H. (2012). The effect of evaluation on teacher performance. *American Economic Review*, 10(7), 3628-3651. doi: 10.1257/ger.102.7.3628

Tracy, S. J. (1995). How historical concepts of supervision relate to supervisory practices today. *The Clearing House*, *68*(5), 320. Retrieved from

http://search.ebscohost.com/login.aspx?direct=true&AuthType=shib&db=edsjsr&AN=edsjsr.30189094&site=eds-live

Turnipseed, S., & Darling-Hammond, L., (2015). Accountability is more than a test score. *Education Policy Analysis Archives, 23*(11). Retrieved from http://search.ebscohost.com/login.aspx?direct=true&AuthType=shib&db=eric&AN=EJ1051725&site=eds-live

U. S. Department of Education. (2009). *Race to the top executive summary* [PDF file]. Retrieved from http://www2.ed.gov/programs/racetothetop/index.html

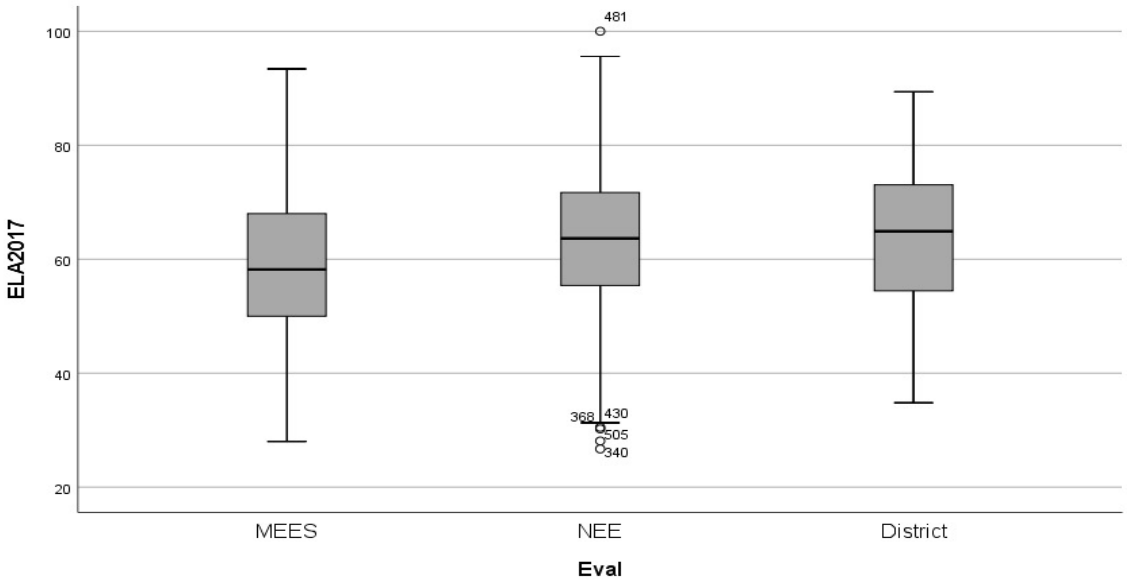
United States National Commission on Excellence in Education. (1983). A nation at risk: the imperative for educational reform: a report to the Nation and the Secretary of Education, United States Department of Education. Washington, D.C.: The Commission: [Supt. of Docs., U.S. G.P.O. distributor].

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

Whitehurst, G. J., Chingos, M. M., & Lindquist, K. M. (2015). Getting classroom observations right. *Education Next, 15*(1), 62-68. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&AuthType=shib&db=eue&AN=99759639&site=eds-live

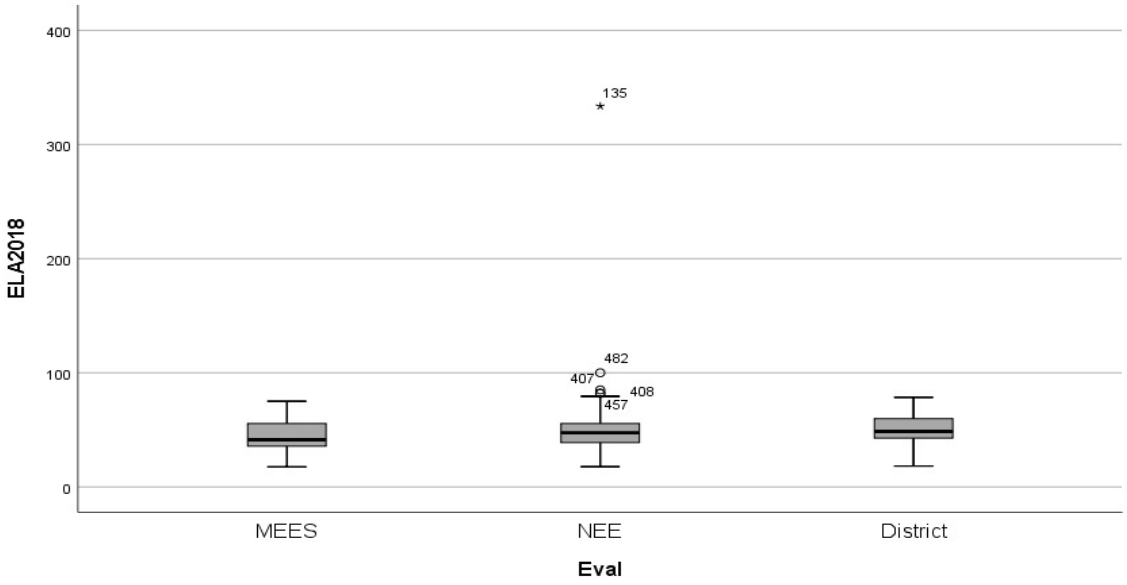
Appendix A

Boxplot of Significant Outliers in Each Evaluation System for the 2017 English Language Arts Missouri Assessment Program



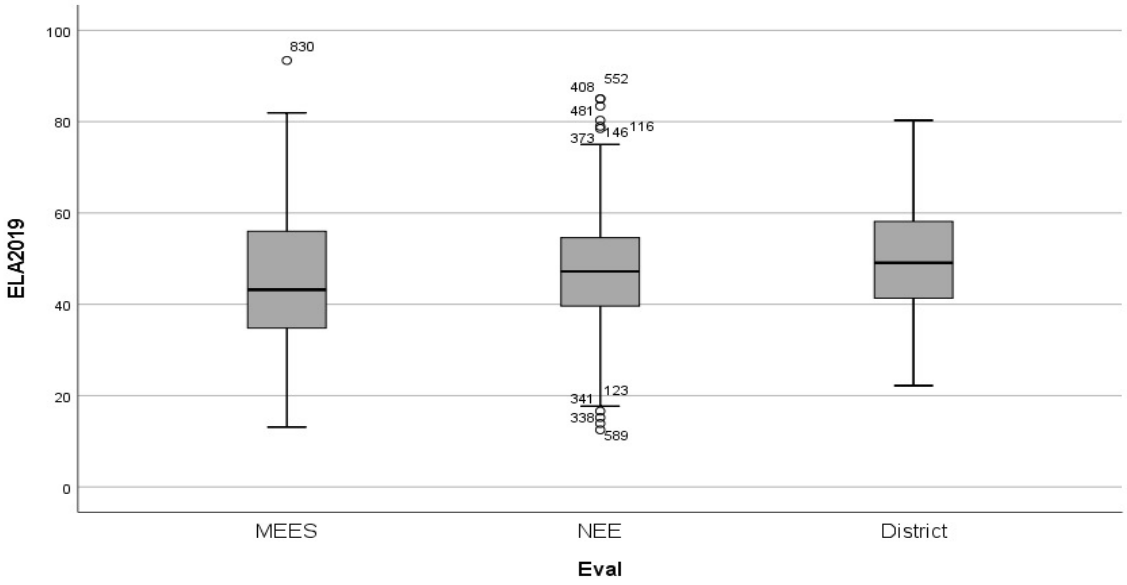
Appendix B

Boxplot of Significant Outliers in Each Evaluation System for the 2018 English Language Arts Missouri Assessment Program



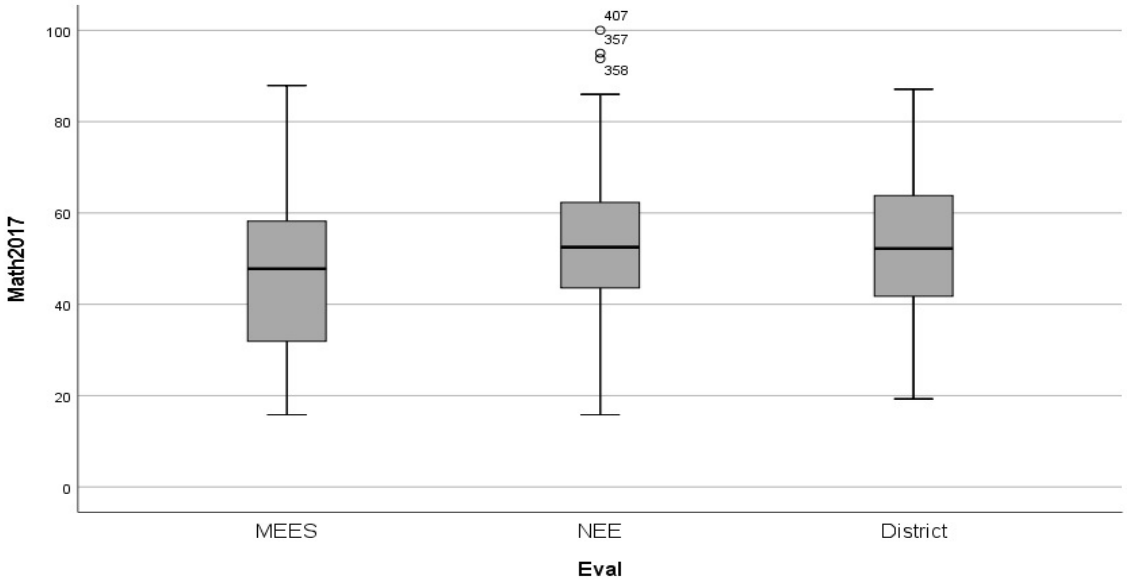
Appendix C

Boxplot of Significant Outliers in Each Evaluation System for the 2019 English Language Arts Missouri Assessment Program



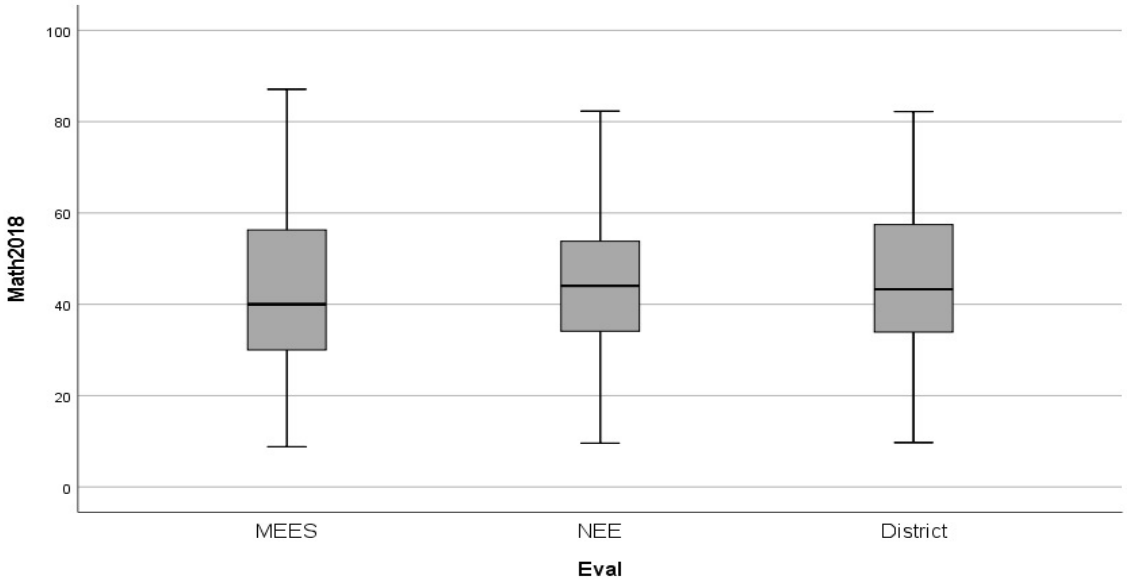
Appendix D

Boxplot of Significant Outliers in Each Evaluation System for the 2017 Math Missouri Assessment Program



Appendix E

Boxplot of Significant Outliers in Each Evaluation System for the 2018 Math Missouri Assessment Program



Appendix F

Boxplot of Significant Outliers in Each Evaluation System for the 2019 Math Missouri Assessment Program