PRINCIPAL AND TEACHER PERCEPTIONS OF THE DURATION OF CLASSROOOM OBSERVATIONS AND EVALUATIVE FEEDBACK

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

PRINCIPAL AND TEACHER PERCEPTIONS OF THE DURATION OF CLASSROOM OBSERVATIONS AND EVALUATIVE FEEDBACK

Presented by David Pyle a candidate for the degree of Doctor of Education and hereby certify that in their opinion it is worthy of acceptance.

Dr. Bill DuVall, Advisor/Chair
Psychology, Southwest Baptist University

Dr. Pamela Hedgpeth, Committee Member
Graduate Education, Southwest Baptist University

Dr. Kathy Tackett, Committee Member
Assistant Superintendent, Carl Junction R-1 School District

PRINCIPAL AND TEACHER PERCEPTIONS OF THE DURATION OF CLASSROOM OBSERVATIONS AND EVALUATIVE FEEDBACK

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

David Pyle, B.S., M.S., Ed.S.

Dr. Bill DuVall, Dissertation Advisor

May 2019

ACKNOWLEDGMENTS

This dissertation is dedicated to my parents, Clyde and Linda Pyle, for teaching me the value of education, perseverance, and resilience. They provided the foundation that allowed me to complete this work. I am also thankful to my wife, Jill, and children, Dylan and Lily, for supporting me throughout this process.

I have appreciated the tremendous professional support, both formally and informally, from my colleagues and Southwest Baptist University professors. Thanks especially to my advisor, Dr. Bill DuVall, and my other committee members, Dr. Pam Hedgpeth and Dr. Kathy Tackett. A full list of those who have supported me is not practical, but I'm especially appreciative of Theresa Wilson, Kyle Williams, Jesse Wall, Dr. Phillip Cook, Dr. Nicole Keller, Jared Wooderson, Christen Glenn, and Dr. Bill Powers.

Support from my friends was a critical part of completing this work. I appreciate those who were willing to listen, especially Steven Taylor and Jonathan Dawson. Finally, thanks to my good friend, Brian, for being a near constant companion during the writing of this dissertation.

TABLE OF CONTENTS

LIST OF TABLES

ABSTRACT

This quantitative study was conducted to describe and compare the perceptions of Missouri secondary principals and teachers regarding the duration of classroom observations and evaluative feedback. The study explored the effects of variation in classroom observation durations from the perspective of both principals and teachers. A survey instrument was used to sample participants to address the research questions guiding the study. Responses from principals indicated greater agreement in perceptions regarding the adequacy of reported observation durations than effects of observation duration. Principals also indicated greater agreement in perceptions regarding the characteristics of evaluation feedback than the impact of the feedback. The perceptions of principals did not differ significantly according to the reported duration of observations. Responses from teachers indicated greater agreement in perceptions regarding the effects of observation duration than the adequacy of reported observation durations. Teachers also indicated greater agreement in perceptions regarding the characteristics of evaluation feedback than the impact of the feedback. The perceptions of teachers reporting shorter observation differed significantly from those reporting longer observation durations. Statistically significant differences were found between the perceptions of principals and teachers. Principals indicated significantly greater agreement in perceptions regarding observation duration and evaluation feedback. The durations and frequencies of observation reported by principals and teachers indicate the trend toward shorter and more frequent observations was present in Missouri secondary schools.

CHAPTER ONE

INTRODUCTION

Teacher evaluation evolved tremendously in structure and philosophy from the early 1980s through the first decade of the 21st century (Campbell, 2013; Danielson & McGreal, 2000; Marshall, 2012). This evolution was influenced by changes in educational theory, criticisms of traditional evaluation systems, and legislative and policy initiatives, such as No Child Left Behind (No Child Left Behind [NCLB], 2002) and Race to the Top (Donaldson, 2016; Marzano, Frontier, & Livingston, 2011; Romano, 2014). Traditional evaluation systems and associated classroom observations did little to differentiate between effective and ineffective teachers or promote professional growth (Colvin, Flannery, Sugai, & Monegan, 2009; Marshall, 2012). No Child Left Behind (2002) waivers have been granted to states that adopted evaluation systems that included measures of student growth and differentiated levels of teacher performance. Race to the Top grants have also incentivized the adoption of these types of evaluation systems. At the same time, educational theory regarding supervision and evaluation has shifted from rigid clinical supervision to flexible teacher development (Marzano et al., 2011). The result has been the adoption of evaluation systems that rely upon multiple measures of effectiveness and focus on increasing student achievement. However, classroom observations remain the most widely utilized tool for teacher evaluation (Cohen & Goldhaber, 2016; Donaldson, 2016).

Educators have increasingly changed their approach to classroom observations by moving to more frequent, shorter mini-observations. This approach provides for greater sampling of instruction over time and more frequent opportunities for evaluators to

provide feedback to teachers (Marshall, 2013). Broad support for the use of feedback to promote professional growth in teachers exists within literature. Increasing the frequency of observation and associated feedback supports professional growth in teachers and improved student achievement (Campbell, 2013; Shana, Glassett, & Copas, 2015; McEntire, 2010; Zamary, 2012).

It is generally accepted that more frequent observation is preferable to less frequent observations (Cohen & Goldhaber, 2016; Donaldson, 2016; Marshall, 2012). However, it remains unclear how many observations provide the desired benefits and at what point additional observations fail to provide additional feedback for growth. Broad variation in the number of observations per teacher has been documented across school districts. Significant questions also exist regarding the duration of classroom observations. As with frequency of observation, broad variation in the duration of classroom observations has been found to occur across school districts (Cohen & Goldhaber, 2016). The amount of time recommended by teacher evaluation experts varies with recommendations ranging from 3 minutes to 10 minutes to 15 minutes and beyond (Downey, Steffy, English, Frase, & Poston, 2004; Marshall, 2013; Marshall & Marshall, 2017). Kitendo (2015) cited the duration of classroom observations as an area in need of additional research.

Theoretical Framework

The theoretical framework for this study spanned disciplines and theorists. Bolman and Deal's (2008) human resource frame emphasized mutually beneficial relationships between individuals and organizations. Bolman and Deal referred to Maslow's hierarchy of needs as well as McGregor's Theory X and Theory Y within the

context of the human resource organizational frame. Maslow described self-actualization as the highest level in the hierarchy. At this level, Maslow (1954) described individuals as maximizing their abilities. McGregor expanded upon Maslow's hierarchy. In Theory Y, he described establishing conditions that allow individuals to become self-directed (McGregor, 1960). Senge (2006) described the discipline of personal mastery, which is increased by individuals intentionally honing their craft through continuous learning. "Reciprocal commitments between individual and organization" (Senge, 2006, p. 8) were emphasized within the discipline of personal mastery. Both individual and organizational benefits are achieved from the promotion of continuous growth. Dweck (2016) supported the interconnectedness of individual and collective efforts in sustaining development through the application of growth mindset. A growth mindset was described as supporting the capacity for ongoing learning and development through individual efforts and with assistance from others.

The concepts of self-actualization, self-directedness, personal mastery, and growth mindset align with professional growth as a primary purpose of teacher evaluation and the theory of developmental supervision. Evaluative feedback has been found to contribute to professional growth when individuals use it for goal setting and to better understand their own practices (Feeney, 2007; Goe, 2013). Developmental models have emphasized teacher development as a primary responsibility of supervision and evaluation. The goal expressed in these models is for teachers to become self-directed in their growth and in improving instruction (Glickman, Gordon, & Ross-Gordon, 2010). Improved instruction has been found to correlate with higher student achievement (Hattie, 2012). Supervisors operate on a continuum of directive, collaborative, and

nondirective behaviors depending upon the level of self-actualization of the teacher. In applying the theory of developmental supervision to teacher evaluation, evaluators have sought to promote the growth of teachers through the evaluation process (Glickman et al., 2010).

The remainder of Chapter One presents the problem statement, the rationale for the study, and the research questions that guided the study. The limitations, delimitations, assumptions, and design controls of the research are also discussed along with key terms of the study.

Problem Statement

There is conflicting guidance about the amount of time that an evaluator needs to remain in a classroom for an observation as well as minimal research on the topic (Kitendo, 2015; Marshall, 2013). Principals have also found it difficult to implement more frequent observations because of time constraints (Donaldson, 2016). This study examined the perceptions of principals and teachers about the duration of classroom observations and evaluative feedback to add to the body of research on the topic.

Variation in practice and recommendations has been found to exist regarding the duration of classroom observations. The duration of classroom observations has been identified as an area for additional research based upon a review of literature and existing research (Cohen & Goldhaber, 2016; Kitendo, 2015; Marshall, 2013). The promotion of the professional growth of teachers to improve student achievement has been identified as a primary purpose of evaluation (Marzano et al., 2011). Classroom observations have been identified as a mechanism for evaluators to provide feedback to teachers. Feedback has been associated with instructional improvement and increased teacher engagement in

professional development (Cohen & Goldhaber, 2016; Donaldson, 2016; Shana et al., 2015).

School districts have increasingly adopted evaluation systems utilizing more frequent mini-observations (Donaldson, 2016). The objective feedback from frequent observations has been associated with positive outcomes, but some aspects remain unclear. It is not understood at what point continued observations provide continued benefits. Variation has also been documented in practice and in what is recommended regarding the duration of classroom observations (Cohen & Goldhaber, 2016).

Principals have been frequently identified as filling the role of instructional leader in schools (Neumerski, 2013). Principals have also been identified as being most commonly assigned the role of evaluator in teacher evaluation systems (Coggshall, Rasmussen, Colton, Milton, & Jacques, 2012). Support has been documented within literature for principals to apply feedback and coaching in evaluative processes to promote teacher growth and increase student achievement (Hattie, 2012; Hattie & Clinton, 2011; McNulty, 2011; Mette et al., 2017).

Rationale for the Study

The purpose of this quantitative study was to describe and compare the perceptions of Missouri secondary principals and teachers regarding the duration of classroom observations and evaluative feedback. Time is a factor both in the duration of observations and in the broader context of time management for school principals. Evaluators, typically school principals, have a variety of responsibilities and demands for their time (Cohen & Goldhaber, 2016; Goe, 2013; Kraft & Gilmour, 2016; Marshall, 2012).

This study provided guidance to evaluators on conducting observations of adequate duration to produce evaluative feedback for professional growth. The study provided guidance for evaluators in minimizing the duration of each classroom observation beyond what is necessary. The study explored principal and teacher perceptions of the duration necessary for a classroom observation to yield high-quality evaluative feedback for professional growth.

Research Questions

The following research questions guided the study:

1. What are the differences in perceptions among secondary principals regarding the duration of classroom observations conducted by principals and evaluative feedback?
2. What are the differences in perceptions among secondary teachers regarding the duration of classroom observations conducted by principals and evaluative feedback?
3. What are the differences between the perceptions of secondary principals and teachers regarding the duration of classroom observations conducted by principals and evaluative feedback?

The following null hypotheses related to the research questions of the study existed:

H_{01}: Significant differences will not exist in the perceptions of secondary principal subgroups regarding the duration of classroom observations conducted by principals and evaluative feedback.

H_{02}: Significant differences will not exist in the perceptions of secondary teacher subgroups regarding the duration of classroom observations conducted by principals and evaluative feedback.

H_{03}: Neither principals nor teachers will perceive that the quality of evaluative feedback is affected by the duration of classroom observations conducted by principals. There will be no significant differences between the perceptions of secondary principals and teachers.

Limitations, Delimitations, and Assumptions

The limitations, delimitations, and assumptions associated with the study are identified in the following sections.

Limitations. The following were limitations of the study.

1. The study was limited by the response rate to the survey instrument utilized in the study.
2. Researcher bias was possible because of direct involvement in teacher evaluation as a principal.
3. Data collection in the fall limited the study by excluding first year teachers.

Delimitations. The following were delimitations of the study.

1. The study included principals and teachers in secondary schools, Grades 9 – 12.
2. The study included public schools in the state of Missouri.
3. The study was limited to the perceptions of principals and teachers as sampled during a specific period.

Assumptions. The following were assumptions of the study.

1. It was assumed that all respondents gave honest responses representative of their professional beliefs.
2. It was assumed that the responses received were representative of the practices and perceptions of secondary principals and teachers throughout the state.

Design Controls

This quantitative study utilized a cross-sectional survey of secondary school principals and teachers in the State of Missouri. The quantitative methods used in the study served as a design control for potential researcher bias. Survey responses were anonymous to encourage honesty by the respondents. Requests for survey responses were made until a sample sufficient for generalizability of the results was achieved.

Respondents were identified as principals or teachers within the survey instrument. Participants were asked to identify the average duration of classroom observations in their settings. Participants also reported the frequency of observation. The remaining survey items used a 5-point Likert scale to measure perceptions about the effects of the duration of classroom observations on the quality and usefulness of evaluative feedback. Survey responses were analyzed to describe the perceptions of participants and to determine if significant differences existed depending upon the duration of classroom observations experienced by the participants. Comparisons were also made between the perceptions of principals and teachers about the effects of the duration of classroom observations and evaluative feedback for professional growth.

Definition of Key Terms

Classroom observation – A classroom visit conducted to contribute to the teacher evaluation process (Danielson & McGreal, 2000).

Feedback – Written or verbal information regarding teaching performance given for improvement (Feeney, 2007).

Professional growth – Change in teaching practices resulting in improved instruction (Danielson & McGreal, 2000).

Mini-observation – A classroom observation that samples a portion of a lesson or class period instead of the entire lesson or class period (Marshall, 2013). Mini-observations are sometimes referred to as walk-through observations. The term mini-observation was used for this study and the term walk-through observation was avoided unless used in quoting another author.

Teacher evaluation – The process of assessing teacher performance to determine effectiveness and contribute to the professional growth of teachers (Marzano et al., 2011).

Summary

This chapter presented an introduction to factors impacting teacher evaluation since the early 1980s. The use of classroom observations as the most common tool for teacher evaluation was presented. The increasing use of more frequent mini-observations was also discussed. The theoretical framework, problem statement, rationale for the study, and research questions were presented. The limitations, delimitations, and assumptions were presented. The design controls were described and key terms were defined. This study explored teacher and principal perceptions about the duration of class observations and evaluative feedback.

Chapter Two presents the review of literature related to this study. The review of literature is arranged from general topics to specific topics. The two general topics are the contemporary history of teacher evaluation and teacher evaluation measures and methods. The contemporary history of teacher evaluation section includes review of the limitations of traditional evaluation, political and policy influences, and developmental models of supervision and evaluation. The teacher evaluation measures and methods section includes the topics of value-added measures, surveys, classroom observations, and feedback and coaching.

Chapter Three describes the methodology of the study. Chapter Four presents the findings of the research. Chapter Five presents a summary of the study and the research results, implications of the results, and recommendations for future research.

Chapter Two

Review of Related Literature

Introduction

Teacher evaluation systems that rely upon multiple measures of effectiveness, such as classroom observations, student learning outcomes, and student surveys, are increasingly utilized in schools today (Cohen & Goldhaber, 2016; Donaldson, 2016). The transition to these types of systems has been influenced by convergent factors including shifts in educational theory, criticisms of traditional evaluation systems, and legislative and policy initiatives (Donaldson, 2016; Marzano et al., 2011; Romano, 2014). This chapter presents the review of literature related to factors influencing teacher evaluation and this study. It is arranged from general topics to specific topics. The two general topics are the contemporary history of teacher evaluation and teacher evaluation measures and methods. The contemporary history of teacher evaluation section includes review of the limitations of traditional evaluation, political and policy influences, and developmental models of supervision and evaluation. The teacher evaluation measures and methods section includes the topics of value-added measures, surveys, classroom observations, and feedback and coaching.

Contemporary History of Teacher Evaluation

The period from the early 1980s through the first decade of the 21st century was a time of transition in teacher evaluation. The focus of teacher evaluation moved from a clinical supervision approach to models focused on teacher development. This transition was influenced by limitations of traditional evaluation systems, political and policy influences, and changes in educational theory and practice. For over 30 years, research

identified flaws in teacher evaluation systems. The prevailing criticism was the failure of teacher evaluation to differentiate between high-performing and low-performing teachers. The research was accompanied by recommendations, such as the use of evaluation systems with multiple performance level descriptors and multiple measures of effectiveness. The United States Department of Education began to incorporate these recommendations into policy through No Child Left Behind (2002) accountability waivers and Race to the Top grants. By 2016, teacher evaluation systems incorporating recommendations for teacher evaluation reforms had been adopted in 46 states. Evaluation continued to be used for high-stakes accountability, including personnel decisions. The emphasis on evaluation as a mechanism for promoting teacher development and student achievement also increased (Donaldson, 2016; Marzano et al., 2011; Romano, 2014).

Limitations of traditional evaluation. Significant studies have documented limitations in teacher evaluation systems from the mid 1980s and into the second decade of the 21st century. One of the most prevailing criticisms throughout the studies was the failure of traditional systems of evaluation to differentiate between effective and ineffective teachers. In addition to documenting flaws in evaluation systems, the studies acted as a catalyst for change (Darling-Hammond, 2013; Marzano & Toth, 2013).

A 1984 Rand Corporation study of effective teacher evaluation practices conducted for the National Institute of Education by Wise, Darling-Hammond, McLaughlin, and Bernstein began the contemporary documentation of limitations in teacher evaluation systems. The study involved the review of teacher evaluation systems in 32 school districts across the United States. This was followed by in-depth case

studies of the evaluation systems of four school districts in Utah, Washington, Connecticut, and Ohio. The districts were selected because of diversity in evaluation procedures as well as organizational and demographic characteristics.

Across the 32 school districts studied, the authors noted that respondents consistently reported that "principals lacked sufficient resolve and competence to evaluate accurately" (Wise et al., 1984, p. 22). Other negative aspects reported included teacher indifference or opposition, inconsistency within school systems, and insufficient evaluator training. Favorable evaluation outcomes reported by the authors were better communication between administrators and teachers and more attention to instructional outcomes and pedagogy by teachers. School districts were more likely to use evaluation results to make personnel decisions about nontenured teachers than tenured teachers.

The authors made five broad conclusions and related recommendations based upon the research. Teacher evaluation systems should align with local characteristics and priorities. School district leaders should support evaluation systems by providing resources, including time, training, and mechanisms for quality assurance. Resource allocation should support reliability, validity, and evaluation outcomes. Teacher evaluation purposes and processes should align. Teachers should be involved in decisions about evaluation systems and processes (Wise et al., 1984).

A report published in 2008 by Toch and Rothman, *Rush to Judgment: Teacher Evaluation in Public Education*, renewed criticism of teacher evaluation systems. The authors characterized evaluation systems in public schools as "superficial, capricious and don't even directly address the quality of instruction, much less measure students' learning" (Toch & Rothman, 2008, p. 1). The report was prepared through review of

research studies including data from the National Council on Teacher Quality and The New Teacher Project.

Citing National Council on Teacher Quality data, the authors stated that one third of the fifty largest school districts in the nation did not require teachers to be evaluated annually (Toch & Rothman, 2008). Twenty-five percent of the districts only required evaluation every third year. The authors were critical of observations because of lack of frequency, poorly trained evaluators, and standards not related to quality instruction. They cited these factors as contributing to evaluation systems that did not produce meaningful feedback and did not differentiate between effective and ineffective teachers. The authors cited New Teacher Project data from the Chicago school district as evidence. The data documented only 0.3% of teachers received unsatisfactory evaluation ratings from 2003–2006 while 93% were evaluated to be at the highest levels.

Toch and Rothman (2008) provided recommendations for improving evaluations. The recommendations included multiple standards-based classroom observations conducted by teams of trained evaluators. The researchers suggested classroom observations be combined with other measures, such as student surveys, student and teacher work samples, and assessment results. Final recommendations included raising the stakes associated with evaluation through pay for performance, sanctions, and tying certification to evaluation (Toch & Rothman, 2008).

Data referenced by Toch and Rothman (2008) were expanded on in a publication from The New Teacher Project the following year. Weisberg, Sexton, Mulhern, and Keeling (2009) used the term "widget effect" to describe "the tendency of school districts to assume classroom effectiveness is the same from teacher to teacher" (p. 4). The

authors cited factors that reinforce the widget effect. These include lack of performance distinctions among teachers and failure to address inadequate performance.

The study was based upon analysis of teacher evaluation systems in 12 school districts: four in Arkansas, two in Colorado, three in Illinois, and three in Ohio. The districts were diverse in size, demographic characteristics, and evaluation systems. The study utilized survey responses from nearly 15,000 teachers and 1,300 administrators within the 12 school districts.

The authors identified five broad factors contributing to the failure of evaluation systems to document and address teacher performance differences. Most teachers, 94% - 99% depending on the number of performance descriptors in the evaluation system, were rated at the highest performance levels with fewer than 1% identified as unsatisfactory. The truly high-performing teachers were not identified and supported. Evaluation was unlikely to lead to professional development. Beginning teachers were not adequately supported. Deficient performances were not identified and addressed through personnel decisions.

The authors claimed that weak procedures and poor application contributed to the ineffectiveness of evaluation systems. The frequency and duration of classroom observations were cited as evidence. Most teachers in the study, 64% tenured and 59% nontenured, were observed two or fewer times annually for an average total of 76 minutes. The average combined annual observation time was 75 minutes for tenured teachers and 81 minutes for nontenured teachers. Deficient performance ratings were not found to lead to increased observation or additional feedback. Sixty-five percent of teachers with high evaluation rankings reported two or fewer annual observations and

62% of those with low evaluation rankings reported the same numbers. Fifty-six percent of teachers with high evaluation rankings as well as 58% of teachers receiving low evaluation rankings reported receiving informal evaluative feedback.

The authors offered four recommendations for improving teacher evaluation systems. Evaluation systems should document differentiated teacher effectiveness in improving student achievement. Evaluators should be trained and held accountable for implementation of evaluation systems. Evaluation results should inform professional learning, pay, and personnel decisions. Evaluation systems should encourage deficient teachers to leave their positions voluntarily, and due process should be just but minimized when dismissal is necessary (Weisberg et al., 2009).

A 2010 publication from the Measures of Effective Teaching (MET) Project, funded by the Bill & Melinda Gates Foundation, cited the widget effect as a primary rationale for studying teacher effectiveness and evaluation systems. The authors of the study contended that "individual teachers receive little feedback on the work they do" and "almost everywhere, teacher evaluation is a perfunctory exercise" (Bill & Melinda Gates Foundation, 2010, p. 3).

The MET study included six urban school districts: Charlotte-Mecklenburg, Dallas, Denver, Hillsborough County, Memphis, and New York City. The study centered around three suppositions. Teacher evaluation should include student achievement measures. Measures other than value-added measures, such as classroom observations and student surveys, should show positive correlations with value-added measures. Evaluation measures should provide specific feedback to promote professional development.

State assessment results for teachers of English language arts and mathematics in Grades 4-8 and high school Algebra I, English Language Arts I and Biology were analyzed in the study. The assessment results were compared to supplemental benchmark assessments in the content areas to compare teacher impact on student performance on both types of assessments. More than 20,000 classroom instruction samples were video recorded as part of the study. Trained evaluators assessed the lessons using five different evaluation rubrics including Charlotte Danielson's framework for teaching. Evidence of teacher pedagogical knowledge, student surveys of instructional perceptions, and teacher surveys of work environment and support were also analyzed in the study.

The study involved an ongoing analysis due to the amount and varied data included. Initial findings of the study supported value-added measures as a predictor of future academic achievement. Variance in inter-year value-added results was noted, but the variance was not great enough to invalidate the use of value-added measures. Positive correlations were found between teacher impact on state assessment achievement and achievement on supplementary assessments. Teacher characteristics as measured by student surveys were found to be stable across groups of students. Student perceptions of effectiveness had positive correlations with academic achievement measures.

The authors of the study provided four recommendations for improving teacher evaluation systems. Teachers should be involved in reviewing lists of students assigned to them for value-added data analysis to promote accuracy. Student surveys should be used across grade levels and content areas to gather student perceptions about instructional effectiveness. Evaluators should receive classroom observation training.

Evaluation data from multiple measures should be accessible to principals and teachers as quickly as possible (Bill & Melinda Gates Foundation, 2010).

Common elements emerged across the studies. All questioned the general validity of evaluations as measures of teacher effectiveness either for poor implementation or failure to differentiate between high and low performance. All recommended evaluator training as a mechanism for improving evaluations (Bill & Melinda Gates Foundation, 2010; Toch & Rothman, 2008; Weisberg et al., 2009; Wise et al., 1984). The three most recent studies recommended coupling classroom observations with other measures, such as student surveys and value-added measures (Bill & Melinda Gates Foundation, 2010; Toch & Rothman, 2008; Weisberg et al., 2009). The recommendations from the studies influenced educational policy and practice through federal government incentives for states to adopt evaluation systems with differentiated levels of performance and measures of student achievement (Marzano & Toth, 2013; The New Teacher Project, 2010).

Despite the changes in evaluation systems, research indicates that evaluations have continued to under-identify teachers in need of improvement. A study of 24 states utilizing evaluation systems with multiple performance levels reported 3% as the median percentage of teachers rated below proficient. However, principals perceived that as many as 28% of teachers performed below proficiency (Kraft & Gilmour, 2017). The degree to which new evaluation systems have been implemented at local levels and improved the effectiveness of teacher evaluation is an area for continued research (Kraft & Gilmour, 2016).

Political and policy influences. The United States Department of Education has influenced teacher evaluation heavily since 2009 through No Child Left Behind (NCLB, 2002) waivers and Race to the Top (RTTT) grants. Forty-six states have implemented evaluation systems that utilize multiple measures of teacher effectiveness (Donaldson, 2016; Romano, 2014). These changes have occurred as states have pursued NCLB (2002) waivers and RTTT grants associated with the adoption of teacher evaluation systems utilizing multiple measures of effectiveness (Aguilar & Richerme, 2014; Croft, Roberts, & Stenhouse, 2016; Polikoff, 2015).

The NCLB Act was passed by Congress in 2001 and signed by President George W. Bush in 2002. The law was an update to the Elementary and Secondary Education Act (ESEA) of 1965. The NCLB (2002) legislation increased accountability for student achievement through standards-based, high-stakes assessment (Duffy, Giordano, Farrell, Paneque, & Crump, 2008). The act set forth a minimum of four performance-level descriptors to define student proficiency on assessments in mathematics and reading. Schools that did not comply or meet proficiency targets risked losing federal education funding (Bailey, 2016; Ellis, 2007; Popham, 2013; Tanner, 2013). The NCLB (2002) system of accountability led to a perception of school failure coupled with motivation for states to set unreasonably low proficiency targets (Croft et al., 2016; Gottlieb, 2013). The prospect of schools failing to meet the mandate of NCLB (2002), all students proficient by 2013–2014, led to the 2011 ESEA Flexibility Program under the administration of President Barack Obama. The ESEA waivers offered states relief from NCLB (2002) requirements in exchange for adopting educational reforms including teacher evaluation and the use of student achievement data in teacher accountability. These were two of the

key components of the Race to the Top program outlined by the U.S. Department of Education under Secretary Arne Duncan in 2009 (Croft et al., 2016; Popham, 2013; Tanner, 2013). Competitive grants available through the RTTT program created an additional incentive for states to apply for NCLB (2002) waivers (Aguilar & Richerme, 2014; Croft et al., 2016). The grants were part of the American Recovery and Reinvestment Act of 2009 and totaled $4,350,000,000. The grants added a weighty financial incentive for states to adopt teacher evaluation systems using multiple measures of effectiveness (LaVenia, Cohen-Vogel, & Lang, 2015; Marzano & Toth, 2013).

President Obama signed the ESSA legislation as a replacement to NCLB (2002) in 2015. The ESSA maintained the test-based accountability of NCLB (2002) but shifted accountability decisions back to the state and district levels. The ESSA also provided for more comprehensive measurement of school performance through both academic and nonacademic measures. Little data has been generated on changes in accountability systems under ESSA (Darrow, 2016; Marsh, Bush-Mecenas, & Hough, 2017).

While the impacts of ESSA have not yet developed, researchers have described the impacts of NCLB and RTTT on evaluation, curriculum, and the standardization of teaching practices. The intent of RTTT was to strengthen the quality of teaching through teacher evaluation systems that differentiate effectiveness with student growth data as a significant factor (Aguilar & Richerme, 2014; Gottlieb, 2013). However, some educators have criticized the emphasis on standardized tests as an evaluative measure. Critics cited concerns about validity and reliability and questioned the viability of test-based accountability as a driver for educational improvement (Aguilar & Richerme, 2014; Croft et al., 2016; Tanner, 2013). Some have claimed that the focus on standardization

diminishes local control of education, narrows curriculum, redefines teaching roles, and limits the pedagogical freedom of teachers through scientific management (Bailey, 2016; Croft et al., 2016; Warring, 2015). In addition to teacher evaluation systems influencing teaching practices, the RTTT grants provided incentives for states to adopt Common Core State Standards (CCSS). Some have called this "coercion" on the part of the federal government to influence the curriculum taught in schools. States aspiring to obtain RTTT funds were more likely to adopt the CCSS and the teacher evaluation requirements associated with the grants (LaVenia et al., 2015).

While the positive or negative impacts of NCLB (2002) and RTTT have been debated, the policy mandates have changed teacher evaluation practices across the nation. The financial incentives and relief from accountability requirements have resulted in most states adopting teacher evaluation systems that utilize multiple performance measures including classroom observations, value-added measures, and surveys (Chin & Goldhaber, 2015; Donaldson, 2016; Marzano & Toth, 2013; Popham, 2013).

Developmental models of supervision and evaluation. Developmental models emphasize teacher growth to improve instruction and increase student achievement as the primary purposes for supervision and evaluation. Knowledge, interpersonal skills, and technical skills have been cited as requirements for supervisors. These elements are applied through directive, collaborative, and nondirective supervisory behaviors. The type of supervisory behaviors utilized vary depending upon the developmental level of the teacher (Glickman et al., 2010). Differentiated approaches are utilized with the goal of teachers becoming self-directed in their professional growth. These models departed from the prescriptive methods of scientific management and clinical supervision. The

use of feedback to promote professional growth is a consistent theme throughout developmental models (Glickman et al., 2010; Marzano et al., 2011).

Developmental models of evaluation evolved from the scientific management and clinical supervision models preceding them. This evolution was influenced by changes in educational theory and criticisms of educational quality. Prior to the 1980s, supervision and evaluation were generalized as ritualistic and authoritarian activities approached in a complacent manner (Sergiovanni & Starratt, 1993).

In the 1983 report, *A Nation at Risk*, the National Commission on Excellence in Education cited a shortage of qualified teachers as a factor negatively impacting the quality of the American educational system. The authors provided a general recommendation for identifying and differentiating between levels of teacher expertise: "School boards, administrators, and teachers should cooperate to develop career ladders for teachers that distinguish among the beginning instructor, the experienced teacher, and the master teacher" (National Commission on Excellence in Education, 1983, p. 31).

A Rand Corporation report prepared for the National Institute of Education was published the following year (Wise et al., 1984). The authors of the report provided criticisms as well as recommendations for improving teacher evaluation. Increased communication between administrators and teachers and an increased focus on instructional outcomes and pedagogy were reported as positive outcomes of evaluation systems. Teacher development was described as a purpose for evaluation. The authors described cooperation, motivation, and efficacy as requisites for evaluation to lead to teacher development (Wise et al., 1984).

The number of states legislating teacher evaluation requirements increased from 26 to 36 within 2 years of the release of *A Nation at Risk* (National Commission on Excellence in Education, 1983) and the Rand report (Wise et al., 1984). While teacher pay incentives were a factor in the increased focus on evaluation, the structure and outcomes of evaluation also changed. Evaluation systems increasingly focused on evaluator training, more frequent formal observation, and feedback for teacher improvement (Brandt, 1995). Throughout the 1980s and 1990s, the concept of instructional leadership increased as a function of supervision and evaluation. The idea that effective instruction could be identified, evaluated, and developed became more common. Frameworks for evaluating instruction and providing feedback were developed and more frequently utilized (Sergiovanni & Starratt, 1993).

Charlotte Danielson created one of the most frequently utilized frameworks for classroom observation. The framework for teaching was originally published in 1996. The Danielson model established rubrics for standards-based evaluation of teachers (Aguilar & Richerme, 2014; Kimball & Milanowski, 2009). The framework included 22 components across four domains. The domains were planning and preparation, classroom environment, instruction, and professional responsibilities. Performance descriptor rubrics were utilized for 66 elements within the components and domains. Four levels of performance, ranging from unsatisfactory to distinguished, were utilized in each rubric. Danielson emphasized collaboration between administrators and teachers and promoted acknowledgement of the characteristics of adult learners and the complexity of teaching. Danielson also supported evaluators understanding the movement of teachers through developmental levels and promoting professional

development (Danielson & McGreal, 2000). Promoting the growth of teachers by understanding adult learning and utilizing differentiated supervisory behaviors continued as a point of emphasis in developmental models (Glickman et al., 2010).

The Danielson framework served as a foundation for other classroom observation models that followed. Robert Marzano's framework for teacher evaluation built upon the concepts of Danielson's framework. The Marzano framework also utilized four domains of teacher performance. The domains were classroom strategies and behaviors, planning and preparing, reflecting on teaching, and collegiality and professionalism. Marzano described five prerequisites for growing expert teachers: articulating the knowledge base for teaching, providing feedback and opportunities for practice, providing opportunities to observe and discuss expertise, establishing clear criteria and planning for success, and recognizing expertise (Marzano et al., 2011). Marzano emphasized the impact of specific instructional methods on student outcomes in support of evaluation as a mechanism for teacher development (Aguilar & Richerme, 2014).

The propagation of standards-based evaluation models has generated some criticism. Critics have contended that teaching is too complex to be standardized. Some have referred to limiting the artistry and creativity of teaching as an unintended consequence of standards-based evaluation (Bailey, 2016; Kimball & Milanowski, 2009). Despite these concerns, the use of standards-based evaluation methods has broadly been supported as the best practice. The use of specific standards measured by rubrics and performance descriptors has increased objectivity in evaluation. It has also supported evaluators in providing specific feedback about teaching (Darling-Hammond, 2013; Marzano & Toth, 2013).

The first decade of the 21st century brought renewed criticisms of teacher evaluation. The authors of *Rush to Judgment* (Toch & Rothman, 2008), *The Widget Effect* (Weisberg et al., 2009), and *Learning About Teaching* (Bill & Melinda Gates Foundation, 2010) were critical of evaluation systems for failing to differentiate levels of performance and effectively promote teacher development. Recommendations from the studies included evaluator training and the use of multiple evaluation measures (Bill & Melinda Gates Foundation, 2010; Toch & Rothman, 2008; Weisberg et al., 2009). These recommendations influenced educational policy and practice through No Child Left Behind (2002) and Race to the Top incentives for states to adopt evaluation systems with differentiated levels of performance and measures of student achievement (Bill & Melinda Gates Foundation, 2010; Marzano & Toth, 2013). By 2016, 46 states had adopted evaluation systems utilizing multiple measures, including student achievement data (Donaldson, 2016). These events established a renewed focus on the use of evaluation as a mechanism for teacher development (Reddy, Dudek, Kettler, Kurz, & Peters, 2016).

Educational reform efforts have increasingly focused on teacher development through evaluation (Cusick, 2014). Research on teacher impact has supported this focus. Of all factors influencing student learning, teachers have been found to have the greatest effect (Hattie, 2012; Lassiter, 2011; Reeves, 2009). Instructional leadership has been established as a secondary factor in student achievement with leaders having the greatest effect by participating in teacher learning and development (Hattie & Clinton, 2011). Formal leaders in schools, principals, have also most commonly filled the role of evaluator. These connections have reinforced the importance of evaluator training in

support of teacher development (Patrick & Mantzicopoulos, 2016). When conducted well, classroom observations have been found to support teacher development and student achievement. When conducted poorly, classroom observations lack validity and reliability; they misidentify teacher performance and fail to produce results (Kimball & Milanowski, 2009). Research has supported the ability of evaluators to provide feedback as an area of emphasis in evaluator training (Kraft & Gilmour, 2016).

Evaluation standards describing specific aspects of teaching and performance descriptors have been cited as the basis for generating evaluative feedback. The usefulness of feedback has been related to various aspects including relevance, accuracy, timeliness, specificity, and immediacy (Reddy et al., 2016; Reeves, 2009). Feedback should be objective, observable, and promote self-reflection (Feeney, 2007). The usefulness of feedback in promoting teacher development is dependent upon characteristics of both the evaluator and the teacher. Evaluators must possess the ability to generate and deliver feedback. The teacher must be willing to receive and act upon the feedback. Evaluator accuracy in generating feedback is influenced by motivation as well as cognitive and contextual factors in assessing instruction. Teachers benefit developmentally from feedback, which supports self-esteem and efficacy (Kimball & Milanowski, 2009).

Teacher Evaluation Measures and Methods

Teacher evaluation systems that utilize multiple measures of teacher effectiveness have been implemented in schools throughout the United States. By 2016, 46 states had adopted evaluation systems utilizing multiple measures. This transition was facilitated by criticism of traditional evaluation systems coupled with federal government policy

influences. Research has supported the use of multiple measures in comprehensive evaluation systems (Donaldson, 2016; Marzano et al., 2011; Romano, 2014).

The most common teacher evaluation measures are value-added assessments, student surveys, and classroom observations (Ferguson & Danielson, 2014). Classroom observations are the most frequently utilized individual measure. Classroom observations are most frequently coupled with value-added measures (Chin & Goldhaber, 2015; Polikoff, 2015). Each measure has documented benefits and limitations that support the use of the measures together.

Classroom observations provide the opportunity for frequent feedback regarding instruction. However, reliability is a concern that increases without clear evaluation rubrics, trained evaluators, and frequent observation (Archer et al., 2016). Observations demonstrate moderate correlations to student achievement increases. Observation results tend to better differentiate between poor and average teachers compared to average and excellent teachers (Polikoff, 2015).

Value-added results have shown the highest correlation to future student achievement (Kane & Staigner, 2012). However, value-added measures exhibit issues with inter-year stability. These measures are impacted by effects not directly attributable to the teacher, such as student demographics and motivation (Warring, 2015). Value-added results also do not provide teachers with specific input for improvement (Ferguson & Danielson, 2014; Raudenbush & Marshall, 2014).

Student surveys are economical to administer and provide input from the individuals most closely engaged in instruction. Sample size and the amount of time engaged with the teacher enhance reliability. Student surveys have most commonly been

used in postsecondary settings. However, the use of student surveys as an evaluation tool in elementary and secondary school settings has increased (Ferguson, 2012).

Value-added measures. Data quantifying changes in student performance across a specified period are referred to as value-added measures (VAMs) or student learning objectives (SLOs). Rather than measuring student achievement at one point in time, growth measures quantify student progress across a period. True value-added measures involve the application of statistical analysis to student achievement data to estimate the effect of the teacher on learning. VAMs compare the expected growth to the actual growth of groups of students with similar characteristics. Teachers with high value-added scores influence student achievement to a greater degree than statistically projected. The complexity of and variability in models have been cited as limiting factors in understanding and applying VAMs (Anderman, Gimbert, O'Connell, & Riegel, 2015; Ballou & Springer, 2015). Limitations associated with VAMs have contributed to greater use of SLOs as growth measures. Student learning objectives compare student growth over time but do not include the application of statistical methods used with VAMs. The SLO method involves pre-assessment, development of growth targets, post-assessment, and comparison of actual growth-to-growth targets (Anderman et al., 2015; Bergin, 2015).

The inclusion of student performance data in evaluation was a requirement of the No Child Left Behind (2002) waivers and Race to the Top grants. A review of state evaluation plans adopted as part of ESEA flexibility or RTTT grant applications showed classroom observations and VAMs as the most common evaluation measures (Polikoff, 2015). Proponents have supported VAMs as an objective method for measuring a

teacher's effect on learning and differentiating between effective and ineffective teachers (Marzano & Toth, 2013; Warring, 2015). Value-added measures have been found to be valuable in measuring student learning when factors such as previous levels of achievement and characteristics of students and schools are considered. The value-added approach has been supported as more valid when comparing the same cohort group instead of different cohort groups at different points in time (Darling-Hammond, 2013; Warring, 2015).

While the use of VAMs has increased the focus on goal-setting and student performance, teachers need additional training on implementing, interpreting, and acting on these measures (Cohen & Goldhaber, 2016; Donaldson, 2016). Questions about the validity, reliability, stability and correlation of VAMs to other measures of teacher quality also exist (Aguilar & Richerme, 2014; Croft et al., 2016; Darling-Hammond, 2013; Polikoff, 2015). Researchers have found significant year-to-year instability in VAMs. Teachers found to be effective based upon VAM results one year could be deemed ineffective the next year and vice versa. These variances were more attributable to differences in student demographics and outside factors than changes in teacher effectiveness (Marzano & Toth, 2013; Warring, 2015). Assessment results have also been found to be impacted by student motivation, a factor not directly attributed to the effectiveness of the teacher (Rutkowski & Wild, 2015). Researchers have found low correlations between VAMs and other teacher evaluation measures, such as classroom observations. Factors influencing this weak correlation include the validity and reliability of each measure, error in one or both measures, and the aspects of teaching captured by each metric (Chin & Goldhaber, 2015; Marzano & Toth, 2013).

In addition to the limitations noted with VAMs, the use of assessments in evaluation has raised ethical issues associated with accountability and responsibility (Beets, 2012; Bolyard, 2015; Kostogriz & Doecke, 2013). Shapiro and Gross (2013) described the ethics of accountability versus responsibility. Accountability is the requirement to demonstrate performance, often as measured by student test scores. Praise or blame is then attributed to schools or individual teachers. In contrast, responsibility is inclusive of society bearing accountability for educating students.

Educational accountability has increased in response to the belief that the public has the right to data reflective of school performance. This movement has been influenced by concerns about the ability of schools to prepare students able to compete in the global economic market (Kostogriz & Doecke, 2013). One response to calls for increased accountability has been incorporation of assessments of student learning into evaluation systems. The intent is to quantify a teacher's impact on student learning (Bolyard, 2015). This type of system makes teachers accountable for student learning outcomes without considering various external factors that impact student achievement (Kostogriz & Doecke, 2013). This has raised concerns about test-based accountability and the issue of accountability versus responsibility.

While the intent of test-based accountability is positive, the practice has generated concerns about the impact on teachers and teaching. Pressure to "teach to the test" at the expense of pedagogy has been noted as an unintended consequence (Beets, 2012; Kostogriz & Doecke, 2013). Curriculum can shrink by focusing on test preparation, and teachers can experience loss of control in instructional decision making (Bailey, 2016; Croft et al., 2016; Hargreaves & Shirley, 2009; Warring, 2015). Test-based evaluation

also does not include stakeholders in sharing accountability for student outcomes. Credit or blame for student achievement is placed primarily on the teacher. External factors, including socioeconomic status and the allocation of resources by stakeholders, are not taken into account (Bolyard, 2015; Kostogriz & Doecke, 2013). This contradiction is central to the issue of shared responsibility. Critics of test-based accountability also contend that the practice interferes with the professional responsibility of teachers to meet the needs of each student. Educators can become more focused on assessment of learning than assessment for learning. Instruction can then become less individualized (Beets, 2012; Kostogriz & Doecke, 2013).

The use of VAMs in evaluation has been controversial, largely because of the limitations associated with VAMs as a measure of teacher quality. Researchers have recommended that VAM results be used with other measures as part of a comprehensive system of evaluation. Evaluators should ensure that sample size is sufficient, should account for demographic factors associated with the school and students, and should examine multiple years of data (Castellano & McCaffrey, 2017; Darling-Hammond, 2013; Marzano & Toth, 2013; Warring, 2015).

Student surveys. Student surveys have been utilized to give teachers feedback as one of the multiple measures in evaluation systems. Surveys are utilized to gather student perceptions about specific aspects of teaching and learning (Danielson & McGreal, 2000; Marzano et al., 2011; Schweig, 2014). The use of student surveys has historically been more common in higher education than primary and secondary education. However, the use of surveys at lower levels has increased as part of evaluation systems utilizing multiple measures (Ferguson, 2012; Polikoff, 2015). Cost

effectiveness and the ability to gather perceptions from individuals directly involved in instruction with teachers for hours each year have been cited as benefits of student surveys (Ferguson, 2012; Van der Lans, Van de Grift, & van Veen, 2015).

Like other evaluation measures, student surveys have been found to have limitations related to reliability with variables such as sample size and the timing of administration affecting the stability of results. Intra-year stability has been found to be higher than inter-year stability because of the effects of different student groups and changes in instruction from year to year (Polikoff, 2015). Outside of these limitations, research has supported the ability of students to identify teacher effectiveness and contribute to formative evaluative feedback (Ferguson, 2012; Van der Lans et al., 2015).

The Tripod student survey was utilized in the Measures of Effective Teaching research. The survey measured student perceptions within seven areas of teacher influence: caring, controlling, clarifying, challenging, captivating, conferring, and consolidating. The study established positive correlations between surveys and VAM scores in mathematics. Students who ranked their teachers in the top 10% through survey responses performed 0.260 standard deviations higher on value-added mathematics measures. In comparison, students who ranked their teachers in the bottom 10% performed 0.244 standard deviations lower on the same assessments (Bill & Melinda Gates Foundation, 2010). Additional analysis of the MET project data found student survey results to be more reliable predictors of VAM results than classroom observations (Kane & Staigner, 2012). These results indicated benefits in coupling student feedback with other observation measures to inform teacher practice and the effect on student achievement (Ferguson, 2012; Marzano & Toth, 2013).

Research into specific domains within the Tripod student survey has indicated that some aspects of student perceptions are better predictors of achievement than others. Students who rated their teachers higher in the areas of controlling and challenging in the Tripod survey achieved higher VAM scores than those who rated their teachers lower in those same areas (Raudenbush & Marshall, 2014). Another study found that control had the highest correlation to student achievement both within the Tripod survey and the Danielson classroom observation framework. These results supported classroom management as a predictor of student achievement (Ferguson & Danielson, 2014).

Classroom observations. Classroom observations remain the most widely utilized tool for teacher evaluation and provide teachers with feedback for setting goals, improving instruction, and advancing student achievement (Cohen & Goldhaber, 2016; Donaldson, 2016; Shana et al., 2015). Classroom observations should provide specific information about teaching practices to be of value to teachers (Colvin et al., 2009). When conducted well, classroom observations have been shown to reveal differentiation in teacher quality (Gottlieb, 2013).

Traditional, scheduled observations have been criticized as "dog and pony shows" that fail to represent the daily instructional practices of the teacher. Marshall (2012) proposed short, unannounced observations as a more effective approach for sampling instruction. Researchers have noted various outcomes associated with frequent mini-observations and feedback (Campbell, 2013; Kitendo, 2015; McEntire, 2010; Shana et al., 2015; Zamary, 2012). Campbell (2013) generalized that teachers and administrators perceive mini-observation models to be preferable to traditional observations due to the frequency and immediacy of feedback. Teachers who are observed more frequently are

more likely to have an increased frequency of participation in professional development (Shana et al., 2015). Teachers are more likely to use the frequent feedback from mini-observations to make changes in instruction (Kitendo, 2015; Zamary, 2012). Research has also shown positive correlations between frequent mini-observations with feedback and improved academic outcomes (McEntire, 2010; Shana et al., 2015).

The objective feedback from frequent observations correlates with positive outcomes, and isolated classroom observations are less likely to present an accurate picture of instruction than more frequent observations. However, some aspects remain unclear. It is not understood at what point continued observations provide continued benefits (Cohen & Goldhaber, 2016). The reliability of classroom observations is enhanced by increasing the number of samples over time. Research has supported a minimum of three to six classroom observations, conducted by at least three evaluators as necessary to provide reliable formative feedback (Polikoff, 2015; Van der Lans et al., 2015; Van der Lans, Van de Grift, van Veen, & Fokkens-Bruinsma, 2016). However, variation has been documented in the actual number of observations conducted from school to school, state to state, and depending upon the tenure status of the teacher. According to the National Center for Education Statistics Schools and Staffing Survey (SSAS) for 2011-2012, the average number of annual observations for nontenured teachers was 3.4 compared to 2.3 for tenured teachers (as cited in Cohen & Goldhaber, 2016).

Researchers reported the duration of classroom observations as another area of variation. The SASS showed mean observation lengths of 45.9 minutes for nontenured teachers and 49.7 minutes for tenured teachers with standard deviations of 22.8 minutes

and 26.9 minutes respectively (as cited in Cohen & Goldhaber, 2016). Kitendo (2015) stated, "There is a serious setback on the issue of classroom walkthroughs because contemporary research does not address the timeframes that work most effectively. As a result, the duration of walkthroughs has varied greatly" (pp. 29-30). The variation in practice and the gap in the research have been compounded by conflicting guidance regarding the amount of time that an evaluator should dedicate to each classroom observation. Recommendations range from as few as 3 minutes in the Downey model to 10 minutes or more (Marshall, 2013; Marshall & Marshall, 2017; Marzano et al., 2011). However, Downey et al. (2004) advocated for the use of frequent, brief classroom visits followed by feedback conversations as a supplement to, not replacement for, formal evaluation. The Measures of Effective Teaching study utilized scoring of 30-minute segments of videotaped lessons. High correlations were found between the scores of individual lesson segments. Data supported the first 30 minutes of a lesson strongly representing the whole lesson (Joe, McClean, & Holtzman, 2014).

Classroom observations have been found to be coupled most frequently with value-added measures in evaluation systems utilizing multiple measures (Chin & Goldhaber, 2015; Patrick & Mantzicopoulos, 2016; Polikoff, 2015). However, research has shown low correlations between the measures (Kraft & Gilmour, 2016; Warring, 2015). This has raised questions about the validity and reliability of each measure and the ability of each measurement to serve as an indicator of teacher quality (Chin & Goldhaber, 2015; Marzano & Toth, 2013). Sources of error in classroom observations attributed to the characteristics of evaluators include the ability to precisely measure instructional quality, bias, and training. Evaluator errors have been shown to skew

observation results. This has raised the concern of poorly implemented observations reflecting evaluator quality instead of teacher quality (Kane & Staigner, 2012; Kraft & Gilmour, 2016).

The validity of classroom observation data has been impacted by the ability of evaluators to assess instruction in a variety of contexts. Feedback generated from classroom observations should be objective and based upon specific criteria for teaching and learning. Evaluators should refer to specific examples to help teachers develop their own understanding of improving instruction (Colvin et al., 2009; Romano, 2014). Rubrics have been recommended as a tool to promote the objectivity and specificity of feedback (Feeney, 2007; Marshall, 2012). Snow (2014) stated that evaluators in high-performing schools preferred evaluation systems that utilized performance rubrics. Additional research has supported the use of rubrics as a mechanism for facilitating common understandings between evaluators and teachers about the aspects of instruction assessed in classroom observations. Rubrics have been implemented extensively in contemporary evaluation systems (Kimball & Milanowski, 2009; Kraft & Gilmour, 2016). Limiting the aspects of instruction assessed per observation has been shown to increase reliability. In a review of MET data, evaluators scored more consistently when using portions of rubrics rather than the entire rubric. This research supported the concept of "cognitive load" as a limiting factor in evaluator accuracy (Joe et al., 2014).

A related component of the cognitive capacity of evaluators, in addition to the number of instructional indicators assessed, are the specific aspects of teaching assessed. Some instructional indicators have been found to be more difficult to assess than others. For example, evaluators tend to demonstrate greater agreement when assessing indicators

related to student behavior than instructional indicators. Other factors impacting the accuracy of observers, such as past experiences, content knowledge, and adherence to observation procedures, have been documented (Kimball & Milanowski, 2009; Bell et al., 2014). These types of factors collectively have been shown to contribute to evaluator bias as a limitation in classroom observations. For example, teachers of students with higher previous achievement levels tend to be rated higher in classroom observations than those with lower previous achievement levels (Warring, 2015). The interjection of factors not directly related to teacher effects into classroom observation ratings has contributed to the concern of misidentifying teacher performance (Murphy & Beretvas, 2015; Park, Chen, & Holtzman, 2014). Evaluator bias has been found to occur both inadvertently and consciously. Principals have reported inflating evaluation ratings for teachers in need of improvement with as many as 25% of teachers performing below proficiency being rated as proficient. Explanations for the inflation of evaluation ratings included limitation of time to support improvement, impact on teacher morale, difficulty associated with personnel changes, and resistance to difficult conversations (Kraft & Gilmour, 2017). Additional research has supported principal development in the ability to engage in difficult conversations as a component of evaluator training (Kraft & Gilmour, 2016; LeFevre & Robinson, 2015).

Evaluator training in the use of evaluation rubrics and protocols has become a point of emphasis. Training has been shown to address sources of evaluator error, increasing the reliability and validity of observations. Recommended components of evaluator training include certification and monitoring of evaluators. The certification and monitoring process should include initial and ongoing agreement with established

thresholds between the trainee and other certified evaluators. Recorded lessons assessed by "master scorers" are often utilized in this process (Bell et al., 2014; Park et al., 2014).

The existing research has supported the use of multiple, independent evaluators to minimize bias and increase the reliability and validity of classroom observations. The reliability of observation assessments has been found to be higher when using multiple observers than using multiple assessments from a lone observer (Polikoff, 2015; Van der Lans et al., 2016). Constraints related to human and capital resources have limited the use of multiple evaluators in schools. The role of classroom observer has most commonly been assigned to school principals. These factors have emphasized the importance of training and support in time management for principals to conduct effective classroom observations (Goe, 2013; Kraft & Gilmour, 2016; Van der Lans et al., 2015).

Feedback and coaching. The feedback generated from teacher evaluation measures has been established as requisite for promoting teacher development and instructional improvement for increasing student achievement (Coggshall et al., 2012). Feedback has been associated with increased teacher participation in professional development and improved academic outcomes for students (Shana et al., 2015; McEntire, 2010). Factors associated with effective feedback include frequency, timeliness, specificity, and objectivity (Campbell, 2013; Kitendo, 2015; Reddy et al., 2016; Zamary, 2012). The usefulness of feedback has been associated with relevance, accuracy, and promotion of self-reflection (Feeney, 2007; Reddy et al., 2016; Reeves, 2009). Objective feedback, based upon specific criteria for teaching and learning, has

been shown to aid teachers in developing increased understanding of instructional improvement (Colvin et al., 2009; Romano, 2014).

Evaluation standards and rubrics with performance descriptors identifying specific aspects of teaching and learning have been widely adopted as mechanisms for generating specific, objective evaluative feedback (Kimball & Milanowski, 2009; Kraft & Gilmour, 2016). Evaluative feedback has been most frequently delivered in post-observation conferences (Coggshall et al., 2012). Classroom observations have been documented as the most frequently utilized tool for teacher evaluation with principals most commonly serving as evaluators (Cohen & Goldhaber, 2016; Donaldson, 2016; Shana et al., 2015). These connections have supported the importance of evaluator training, including the ability to deliver feedback, in support of teacher development and teacher efficacy (Kimball & Milanowski, 2009; Kraft & Gilmour, 2016; Patrick & Mantzicopoulos, 2016).

Garmston and Wellman (2009) described efficacy as an individual's belief in the capacity and willingness to effect change. Feedback and coaching have served as mechanisms to promote reflection, support efficacy, and promote teacher growth (Costa & Garmston, 2015). Stone and Heen (2014) defined the purpose of evaluative feedback as "to rate or rank against a set of standards, to align expectations, to inform decision making" and coaching feedback as "to expand knowledge, sharpen skill and improve capability" (p. 35). Some have contended coaching and evaluation are separate functions while others have supported coaching as a function of evaluation (Mette et al., 2017; Neumerski, 2013; Reeves, 2009). Bolman and Deal (2008) described empowerment as a leadership characteristic and coaching as a mechanism for increasing the competence of

individuals with high commitment. Coaching has been identified as a mechanism for using feedback to increase teacher efficacy through identification of current performance coupled with opportunities to practice and improve (Hattie, 2012; McNulty, 2011; Reeves, 2009). Costa and Garmston (2015) described "calibrating conversations" as a mechanism with evaluative and coaching components for teachers to use rubrics to identify current performance and to set specific goals for growth. Using formative processes to promote teacher growth has been documented as a key factor in distinguishing coaching from summative evaluation (Mette et al., 2017).

The development of trusting relationships, effective communication, and promotion of self-reflection have been identified as elements of effective coaching. Trusting relationships between the coach and coached individual have been described as supportive of effective two-way communication (McDowall, Freeman, & Marshall, 2014; Mette et al., 2017; Walkowiak, 2016). Costa and Garmston (2016) described coaches as mediators of thinking through reflective conversations. Effective coaching conversations have been described as promoting self-reflection by the teacher regarding specific evidence of teaching and learning followed by goal-setting for improved performance (Costa & Garmston, 2015; Walkowiak, 2016). Coaching has been associated with improved student achievement and increased teacher perceptions of the ability to impact achievement (Akhavan & Tracz, 2016). Coaching has also been associated with both increased initial implementation of strategies at the classroom level and the sustained implementation of strategies (Reinke, Stormont, Herman, & Newcomer, 2014).

Principals have been frequently identified as the instructional leaders in schools (Neumerski, 2013). Additionally, principals have been identified as most frequently

filling the role of evaluator in teacher evaluation systems (Coggshall et al., 2012). Given the roles of principals as evaluators and instructional leaders, support has been documented within literature for the application of feedback and coaching in evaluative processes to promote teacher growth and increase student achievement (Hattie, 2012; Hattie & Clinton, 2011; McNulty, 2011; Mette et al., 2017).

Summary

This chapter presented the review of literature related to this study. Teacher evaluation systems that rely upon multiple measures of effectiveness, such as classroom observations, student learning outcomes, and student surveys, are increasingly utilized in schools today (Cohen & Goldhaber, 2016; Donaldson, 2016). The transition to these types of systems has been influenced by convergent factors including shifts in educational theory, criticisms of traditional evaluation systems, and legislative and policy initiatives (Donaldson, 2016; Marzano et al., 2011; Romano, 2014).

The review of literature was arranged by topic. The two general topics were the contemporary history of teacher evaluation and teacher evaluation measures and methods. The contemporary history of teacher evaluation section included review of the limitations of traditional evaluation, political and policy influences, and developmental models of supervision and evaluation. The teacher evaluation measures and methods section included the topics of value-added measures, surveys, classroom observations, and feedback and coaching.

Chapter Three describes the methodology of the study including the participants and research design, research questions and hypotheses, and research procedures utilized in the study. Chapter Four presents the findings of the research. Chapter Five presents a

summary of the study and the research results, implications of the results, and recommendations for future research.

Chapter Three

Research Design and Methodology

Introduction

The purpose of this quantitative study was to describe and compare the perceptions of Missouri secondary principals and teachers on the effects of the duration of classroom observations on evaluative feedback. School districts have increasingly adopted evaluation systems that rely on more frequent and shorter classroom observations (Donaldson, 2016). The study explored the effects of variation in classroom observation durations from the perspective of both principals and teachers.

A cross-sectional survey was used to sample participants. Surveys were e-mailed to high school principals with a request that the principals forward the survey to the teachers in their schools. Participants identified themselves as principals or teachers and described the average duration of classroom observations in their settings. Participants also reported the frequency of observation. The remaining survey items gathered information about principal and teacher perceptions regarding the effect of the duration of observation on evaluative feedback. This chapter describes the research questions and hypotheses, the participants and research design, and research procedures utilized in the study.

Research Questions and Hypotheses

The following research questions guided the study:

1. What are the differences in perceptions among secondary principals regarding the duration of classroom observations conducted by principals and evaluative feedback?

2. What are the differences in perceptions among secondary teachers regarding the duration of classroom observations conducted by principals and evaluative feedback?
3. What are differences between the perceptions of secondary principals and teachers regarding the duration of classroom observations conducted by principals and evaluative feedback?

The following null hypotheses related to the research questions of the study existed:

H_{01}: Significant differences will not exist in the perceptions of secondary principal subgroups regarding the duration of classroom observations conducted by principals and evaluative feedback.

H_{02}: Significant differences will not exist in the perceptions of secondary teacher subgroups regarding the duration of classroom observations conducted by principals and evaluative feedback.

H_{03}: Neither principals nor teachers will perceive that the quality of evaluative feedback is affected by the duration of classroom observations conducted by principals. There will be no significant differences between the perceptions of secondary principals and teachers.

Research Design and Participants

The study utilized a purposive sample of secondary school principals and teachers. The study was limited to educators in the state of Missouri. There were 550 high school principals and 20,188 high school teachers according to the 2018 Missouri Department of Elementary and Secondary Education (DESE) *2016-2017 Statistics of*

Missouri Public Schools. A list of secondary school principals was obtained from the school directory available on the DESE website. The directory included names and e-mail addresses for the secondary principals. The survey instrument was distributed by e-mail to Missouri high school principals with a request to forward the survey to teachers under their supervision.

Questionnaire. The study utilized a survey instrument that paired a survey construct developed by the researcher with a construct from an existing instrument, the Teacher Evaluation Experience Scale. Permission to use the scale was obtained from the authors. This approach was utilized because an existing instrument, which fully addressed the research questions of the study, was not available. An evaluation feedback construct from the existing instrument was incorporated with a new duration of classroom observations construct. The survey included a demographic and background information section. The remaining items related to the two constructs, Duration of Classroom Observations and Evaluation Feedback. See Appendix A for the full survey. The demographic and background information section of the survey was developed to create groups and subgroups for comparative purposes and to report descriptive aspects of classroom observation practices. Respondents identified themselves as principals or teachers. The demographic and background section also provided descriptive data regarding duration and frequency of classroom observations. These elements relate to the validity, reliability, and impact of classroom observations as an evaluative practice (Cohen & Goldhaber, 2016).

The Duration of Classroom Observations items were developed based upon researcher expertise and review of literature. The items were designed to gather

perceptive data about beliefs, experiences, and preferences regarding the duration of classroom observations. A 5-point Likert scale ranging from *strongly disagree (*1*)* to *strongly agree (*5*)* was utilized with these 10 items. Participants responded to statements related to their beliefs about the impact of the duration of classroom observations on evaluative feedback and the validity of observations as a measure of instructional effectiveness. Variation was noted both in what is recommended and what is practiced in the duration of classroom observations (Cohen & Goldhaber, 2016; Marshall, 2013; Marzano et al., 2011). The purpose of these items was to measure the perceptions of the participants related to the effects of these variations.

The first three items in the duration of classroom observations construct used the question stem "The duration of my classroom observations affected." The items measured perceptions about the impact of observation duration on the quality and usefulness of feedback and validity of measuring instructional effectiveness. The next four items used the question stem "My classroom observations were long enough to." These items measured perceptions about the impact of observation duration on the quality and usefulness of feedback and validity of measuring instructional effectiveness based upon the experiences of respondents. The next item was "Part of an observed lesson represented the quality of the whole lesson." Research has shown high correlations between the scores of individual lesson segments with data supporting the first thirty minutes of a lesson strongly representing the whole lesson (Joe et al., 2014). The last two items were "I would have preferred for my classroom observations to be longer in duration" and "I would have preferred shorter observations over longer observations."

These items measured the preferences of respondents regarding the duration of observation.

The items related to evaluation feedback gathered perceptive data about the quality and impact of evaluative feedback. A 5-point Likert scale ranging from *strongly disagree (*1*)* to *strongly agree (*5*)* was utilized with these 13 items. Participants responded to statements related to their perceptions about the timeliness and objectivity of feedback in their experiences. Participants also responded about the usefulness of feedback for seeking professional development and instructional planning. Timely, objective, and descriptive feedback from classroom observations has been noted as an element influencing the professional growth of teachers (Donaldson, 2016; Feeney, 2007). The purpose of these items was to measure perceptions about the quality and usefulness of evaluative feedback.

The researcher used the Evaluation Feedback construct items from the Teacher Evaluation Experiences Survey – Teacher Form (TEES-T) and the Teacher Evaluation Experiences Survey – Principal Form (TEES-P). The surveys were developed by Linda Reddy, Christopher Dudek, Ryan Kettler, Alexander Kurz, and Stephanie Peters of Rutgers University. On behalf of all authors, Dr. Linda Reddy gave permission to use the scales via e-mail, attached in Appendix B. The full instruments included four constructs, evaluation system, evaluation feedback, evaluation process, and motivation to change. The researcher chose to use the evaluation feedback construct of the TEES-T and TEES-P and pair it with a duration of classroom observations construct to best address the research questions of the study. The other constructs were not used in this study.

The evaluation feedback construct of the TEES-T and TEES-P was developed based upon literature review and expertise of the authors. The authors stated the instruments were intended "to capture educators' attitudes and beliefs about teacher evaluation, operating under the assumption that teacher evaluation should measure and promote effective teaching" (Reddy et al., 2016, p. 122). See Appendix C for the TEES-T and TEES-P evaluation feedback construct. The TEES-T and TEES-P contained fifteen items in the evaluation feedback construct. However, the authors eliminated Items 8 and 9 during reliability and validity testing. Good reliability ($\alpha = .95$) was reported for the evaluation feedback construct of the TEES-T (Reddy et al., 2016).

The researcher used the remaining 13 items of the TEES-T and TEES-P in a branching QuestionPro survey. The survey branched to the TEES-T feedback construct items for respondents who identified as a teacher. The survey branched to the TEES-P feedback construct items for respondents who identified as a principal. The question stem of "The evaluation feedback" was used with the first 12 TEES-T items. The last TEES-T item used the question stem "I was satisfied with the feedback I received." The question stem of "my evaluation feedback" was used with the first 12 TEES-P items. The last TEES-P item used the question stem "I was satisfied with the feedback I provided." The items gathered teacher perceptions about evaluation feedback received and principal perceptions about evaluation feedback provided (see Appendix A).

Validity and reliability were established for the Duration of Classroom Observations construct. Validity and reliability were not reestablished for the Evaluation Feedback construct of the TEES-T and TEES-P. Validity and reliability ($\alpha = .95$) of the TEES-T had already been established and reported by the authors (Reddy et al., 2016).

Consent. The survey instrument was e-mailed to 550 secondary principals with an informed consent statement and a request to forward the survey to teachers under their supervision. The consent statement informed participants that completion of the survey was voluntary and could be discontinued at any time without penalty. Participants were informed that responses were not required for all items and the anonymous results would be presented in summary form.

Selection/sampling. Survey responses were collected through QuestionPro. Responses were anonymous and included no personally identifiable information and no information specific to any school or school district. It was assumed that participants provided honest responses representative of their personal beliefs. Biases could exist based upon the personal experiences of participants, such as positive or negative evaluations.

Requests for participation were sent by e-mail to 550 secondary principals with a request to forward the survey to teachers under their supervision. A two-week response window was planned. A reminder e-mail was sent at the end of the first week. The survey window remained open for a third week due to continued responses from principals and teachers. Out of the 550 principals, 195 gave consent and forwarded the survey on to teachers under their supervision. Survey responses were received from 195 principals and 498 teachers. The response rate was representative of 35.5% of the high school principal population and 2.5% of the high school teacher population.

Instrumentation

In accordance with the guidelines of Southwest Baptist University regarding the protection of human participants, a request for review was submitted to the Research

Review Board (RRB) for approval to survey a sample of 550 principals and 20,188 teachers for this study. After receiving RRB approval on May 30, 2018, participant recruitment and data collection began. The pilot survey was distributed to junior high and middle level principals and teachers between August 7 and August 21, 2018. Pilot survey data were used to establish validity and reliability.

Validity and Reliability

The survey instrument was evaluated for face and content validity prior to distribution for pilot testing. Face validity was established at the researcher designed the survey items. Content validity was established using an index of item objective congruence.

Face validity. The researcher conducted face validity analysis when writing survey items to determine if the items addressed the research questions guiding the study. A panel of experts reviewed the items for face validity. Survey items were determined to address the research questions based upon face validity.

Content validity. A panel of teacher evaluation experts established content validity of the survey items using an index of item objective congruence assessment. Five experts reviewed the survey items. Positions held by the five experts included two assistant superintendents for curriculum and instruction, one middle level principal, one high school teacher, and a data analysis and evaluation specialist for the University of Missouri Network for Educator Effectiveness. Three of the experts held doctoral degrees.

The survey was divided into three sections: Demographics and Background Information, Duration of Classroom Observations, and Evaluation Feedback. The

experts were asked to evaluate the Duration of Classroom Observations items using a 3-point Item Objective Congruency scale: *Good Match* (1), Neutral (0), *Does Not Match* (-1; Rovinelli & Hambleton, 1977). Nine of the 10 survey items were rated as a *Good Match* (1) by all five experts. The third item in the construct was rated as a *Good Match* (1) by four experts and Neutral (0) by one expert, resulting in an average rating of .80. The cut score for item removal or revision was .67. No items were assessed below .67. No amendments were made to the survey items, and all items were retained for pilot testing. Table 1 contains the table of specifications for the pilot survey and an Index of Item Objective Congruence for each item based upon expert evaluation.

Table 1

Table of Specifications and Index of Item Objective Congruence for Duration of Classroom Observations Construct Items

Item	Rating
1. The duration of my classroom observations affected the usefulness of evaluative feedback.	1
2. The duration of my classroom observations affected the quality of the evaluative feedback.	1
3. The duration of my classroom observations affected the validity of the observations.	.8
4. My classroom observations were long enough to assess instructional effectiveness.	1
5. My classroom observations were long enough to provide useful feedback.	1
6. My classroom observations were long enough to provide quality feedback.	1
7. My classroom observations were long enough to accurately reflect instructional practices.	1
8. Part of an observed lesson represented the quality of the whole lesson.	1
9. I would have preferred for my classroom observations to be longer in duration.	1
10. I would have preferred shorter observations over longer observations.	1

Pilot

The pilot survey was distributed to junior high and middle level principals and teachers between August 7 and August 21, 2018. Survey responses were received from 43 principals and 49 teachers. Individual scores for the pilot survey items were in the form of Likert scale responses for each item. The responses ranged from 1 (*strongly disagree*) to 5 (*strongly agree*). The demographic information collected in the survey was used for grouping of participants. Survey responses were exported from QuestionPro into Statistical Package for the Social Sciences (SPSS) software. The SPSS software was utilized to conduct exploratory factor analysis of survey items to determine construct validity. Cronbach's alpha was used to establish reliability. Amendments were made to the survey based upon reliability and validity measures from the pilot survey data.

Construct validity. Pilot survey items were analyzed to establish validity and reliability. The Duration of Classroom Observation scale was evaluated for validity by exploratory factor analysis of the survey scale. The scale was analyzed with SPSS software using the extraction method of Principal Component Analysis. Two components with eigenvalues greater than 1 (actual eigenvalues = 3.5 and 2.7) were extracted from the principal survey responses. Three components with eigenvalues of 1 or greater (actual eigenvalues = 4.6, 2.1 and 1.0) were extracted from the teacher survey responses.

Factor analysis was utilized to determine if survey items measured the intended construct and functioned together. Exploratory factor analysis results are displayed in Appendixes D and E. Values of .30 or greater are included.

The Duration of Classroom Observations survey items appeared to measure a single construct based on face and content validity. However, Items 1–3 loaded as one component and Items 4–7 loaded as a second component for both principal and teacher responses. The researcher determined that the wording of the items affected the loading as two separate components. The wording of survey items can result in the extraction of more than one component while the items still measure the same construct (Hof, 2012). Items 1-3 measured general perceptions regarding effects of observation duration. Items 4-7 were more specific to the length of observation experienced by respondents. The researcher determined that Items 1-7 best addressed the research questions of the study. Items 8-10 were eliminated because the items did not show strong alignment with Items 1-7.

Factor analysis using the extraction method of Principal Component Analysis was repeated for Items 1-7 only. The items loaded as two components. The eigenvalues for the components were 2.7 and 2.5 for principal responses. The components accounted for 74.4% of variance. The eigenvalues for the components were 3.4 and 2.1 for teacher responses. The components accounted for 78.5% of variance. Appendixes F and G display the component matrices and scree plots for items 1-7. Values of .30 or greater are included.

Reliability. Items 1-7 were evaluated for reliability. Cronbach's alpha was used to assess reliability and to determine the internal consistency of the items. Cronbach's alpha scores near 1 show high reliability of survey items. The items demonstrated acceptable internal consistency. The reliability for principal responses (n = 43) was α = .73. The reliability for teacher responses (n = 48) was α = .68.

Summary of reliability and validity. The preliminary Table of Specifications (Table 1) showed alignment in the Duration of Classroom Observations scale. Expert analysis established content validity in the scale. Exploratory factor analysis was used to establish construct validity and resulted in the elimination of Items 8-10 in the Duration of Classroom Observations scale from the final survey. Cronbach's alpha established reliability in the remaining items.

Final Survey

The researcher used face, content, and construct validity and Cronbach's alpha reliability analysis to make survey revisions. The survey disseminated to principals and teachers for data collection is in Appendix H. The survey was distributed by e-mail to 550 secondary principals with an informed consent statement and a request to forward the survey to teachers under their supervision (see Appendix I). Survey distribution began on September 4, 2018. A reminder message was sent at the end of the first week. Participants were allowed 3 weeks to respond. Data collection ended on September 25, 2018. The distribution of surveys avoided significant school events, such as breaks and state assessments, to minimize conflicts that could limit participation. Survey responses were received from 195 principals and 498 teachers.

Validity. The survey instrument contained two constructs. SPSS software was utilized to repeat exploratory factor analysis of Duration of Classroom Observations construct items to verify construct validity. Factor analysis was utilized to determine if survey items measured the intended construct and functioned together.

Factor analysis using the extraction method of Principal Component Analysis was repeated for the final survey items in the Duration of Classroom Observations construct.

The items loaded as two components. The eigenvalues for the components were 3.2 and 2.4 for principal responses. The components accounted for 80.5% of variance. The eigenvalues for the components were 3.3 and 2.5 for teacher responses. The components accounted for 83.2% of variance. The first three survey items loaded together as one component and the remaining four survey items loaded together as another component. Tables 2 and 3 display the component matrices for the final survey items. Values of .60 or greater are included. See Appendix J for scree plots.

Table 2

Rotated Component Matrix for Factor Loadings With Exploratory Factor Analysis With Principal Component Analysis (Final Survey Principal Responses)

Item	Component 1	Component 2
1. The duration of my classroom observations affected the usefulness of evaluative feedback.		.94
2. The duration of my classroom observations affected the quality of the evaluative feedback.		.93
3. The duration of my classroom observations affected the validity of the observations.		.87
4. My classroom observations were long enough to assess instructional effectiveness.	.88	
5. My classroom observations were long enough to provide useful feedback.	.91	
6. My classroom observations were long enough to provide quality feedback.	.90	
7. My classroom observations were long enough to accurately reflect instructional practices.	.84	

Reliability. The Duration of Classroom Observations items in the final survey were evaluated for reliability. Cronbach's alpha was used to assess reliability and to determine the internal consistency of the items. Cronbach's alpha scores near 1 show high reliability of survey items. The items demonstrated acceptable internal consistency.

The reliability for principal responses (n = 190) was α = .77. The reliability for teacher responses (n = 489) was α = .79. The two components extracted from the construct were also evaluated for reliability. The researcher reversed the component numbers (1 and 2)

Table 3

Rotated Component Matrix for Factor Loadings with Exploratory Factor Analysis with Principal Component Analysis (Final Survey Teacher Responses)

Item	Component 1	Component 2
1. The duration of my classroom observations affected the usefulness of evaluative feedback.		.91
2. The duration of my classroom observations affected the quality of the evaluative feedback.		.93
3. The duration of my classroom observations affected the validity of the observations.		.89
4. My classroom observations were long enough to assess instructional effectiveness.	.92	
5. My classroom observations were long enough to provide useful feedback.	.92	
6. My classroom observations were long enough to provide quality feedback.	.93	
7. My classroom observations were long enough to accurately reflect instructional practices.	.80	

to align with the numbering sequence of the survey items. The reliability for Component 1 (Items 1-3) for principal responses (n = 191) was α = .90. The reliability for Component 1 (Items 1-3) for teacher responses (n = 493) was α = .90. The reliability for Component 2 (Items 4-7) for principal responses (n = 190) was α = .90. The reliability for Component 2 (Items 4-7) for teacher responses (n = 492) was α = .93. The items in both components demonstrated excellent internal consistency.

Research Procedures

A survey instrument was developed and distributed to Missouri secondary principals and teachers for completion. Survey responses were collected in QuestionPro and imported to SPSS for analysis. Exploratory factor analysis and reliability testing was repeated. All survey items were retained for data analysis. Descriptive statistics of mean and standard deviation were calculated for each valid and reliable survey item for both participant groups, principals and teachers. Inferential statistical analysis was completed for survey constructs and individual items to generalize responses from the high school principal and teacher populations. A one-way analysis of variance (ANOVA) was used to compare the means of survey responses and test for statistical significance between groups. Significance among groups was tested by Tukey's honest significant difference (HSD). The primary groups, principals and teachers, were subgrouped based upon the average duration of classroom observations reported by the participants.

Means for each survey item related to the Duration of Classroom Observations construct were calculated for principals and teachers. The mean for each item was reported to compare the perceptions of principals and teachers within this construct. The mean and standard deviation for all items contributing to the construct were also calculated for each group. Three subgroups were established based upon the average duration of classroom observations reported by the participants in each group. The subgroups were based upon observation durations of less than 15 minutes, 15-30 minutes, and more than 30 minutes. It was assumed survey response would provide sufficient sample sizes in each group. The subgroups would have been adjusted if any group lacked a sufficient sample size. A one-way ANOVA was used to determine statistical

significance. Tukey's HSD was used post hoc to verify significance differences among observation duration subgroups within the teacher and principal groups. The data from these tests were used to analyze the first and second research questions.

Means for the survey items related to the Evaluation Feedback construct were calculated for principals and teachers. The mean for each item was reported to compare the perceptions of principals and teachers within this construct. The mean and standard deviation for all items contributing to the construct were also calculated for each group. A one-way ANOVA was used to determine statistical significance. Tukey's HSD was used post hoc to verify significance among observation duration subgroups. The data from these tests were used to analyze the first and second research questions.

A 3 x 2 factorial design was used to determine statistical significance among the observation duration subgroups within the principal and teacher groups. Tukey's HSD was used to test for significant differences. The analysis was conducted for both survey constructs. Educational role, principal or teacher, served as one grouping in the factorial design. The other grouping was based upon reported duration of observation: less than 15 minutes, 15-30 minutes, and more than 30 minutes. It was assumed survey response would provide sufficient sample sizes in each group. The subgroups and method of statistical analysis would have been adjusted had any group lacked a sufficient sample size. Comparisons were made between the principal and teacher groups and subgroups using the inferential statistics for the two survey constructs to analyze the third research question. The frequency of observation responses from the survey were reported for the principal and teacher groups. These data were used in analyzing and interpreting factors that could influence the perceptions of the participants.

Summary

This chapter presented the methodology of the study including the research questions and hypotheses, research design and participants, and research procedures. The study used a survey instrument that paired a newly developed construct with a construct from an existing instrument. The survey was tested for validity and reliability and used to measure the perceptions of secondary principals and teachers about the duration of classroom observations and the quality of evaluative feedback. Descriptive statistics for survey items were calculated. Participants were subgrouped by the average duration of classroom observations reported. One-way ANOVA and factorial design analysis were used to test for statistically significant differences within the groups and subgroups. Tukey's HSD was used to test for significance among the groups and subgroups. Results from the tests were utilized to compare the perceptions of principals and teachers.

The following chapters present the results obtained from application of the research method and an analysis of the implications of the results. Chapter Four presents the findings of the research. Chapter Five presents a summary of the study and the research results, implications of the results, and recommendations for future research.

Chapter Four

Analysis of the Data

Introduction

This quantitative study was conducted to describe and compare the perceptions of Missouri secondary principals and teachers regarding the duration of classroom observations and evaluative feedback. The study explored the effects of variation in classroom observation durations from the perspective of both principals and teachers. A survey instrument was used to sample participants to address the research questions guiding the study.

Three survey items composed a demographic and background information section. The items were used to create groups and subgroups for comparative purposes and to describe evaluation practices. Respondents identified themselves as principals or teachers. The remaining two items in this section gathered descriptive data regarding duration and frequency of classroom observations. The 20 additional survey items measured perceptions regarding the duration of classroom observations and evaluation feedback.

Survey responses for principals and teachers were downloaded from QuestionPro to separate SPSS data sets. All responses were copied into a third combined data set. All survey responses were retained for data analysis. The responses for all items contributing to each survey construct were combined to create aggregate variables for the constructs. Data cleaning was limited to the default SPSS software settings to exclude missing data. Descriptive statistics utilized pairwise exclusion of missing values. Inferential statistics utilized listwise exclusion of missing values.

Assumption testing for ANOVA included normality and homoscedasticity. Residuals were calculated for survey responses and tested for normality. The residuals for the Duration of Classroom Observations construct were non-normally distributed with skewness of -.66 (*SE* = .09) and kurtosis of 2.55 (*SE* = .19). The residuals for the Evaluation Feedback (TEES-T and TEES-P) construct were non-normally distributed with skewness of -.62 (*SE* = .10) and kurtosis of 1.5 (*SE* = .199). Welch and Brown-Forsythe tests of homoscedasticity of the residuals for both constructs did not indicate significance. Despite the non-normal distribution of data, the researcher proceeded with the use of ANOVA to compare means because skew and kurtosis were within acceptable limits, and homogeneity of variance was assumed. The Kruskal-Wallis test was used as a nonparametric test of significant differences to support ANOVA results.

Descriptive statistics of mean and standard deviation were calculated for the seven Duration of Classroom Observations construct items and the 13 Evaluation Feedback (TEES-T and TEES-P) construct items for both participant groups, principals and teachers. Inferential statistical analysis was completed for both survey constructs and individual items to generalize responses from the high school principal and teacher populations. A one-way ANOVA was used to compare the means of survey responses and test for statistical significance between groups. Significance among groups was tested by Tukey's HSD. The primary groups, principals and teachers, were subgrouped based upon the average duration of classroom observations reported by the participants.

Responses for survey items related to the Duration of Classroom Observations construct were analyzed for principals and teachers. The mean for each item was reported to describe the perceptions of principals and teachers within this construct. The

mean and standard deviation for all items contributing to the construct were also calculated for each group. Three subgroups were established based upon the average duration of classroom observations reported by the participants in each group. The subgroups were based upon observation durations of less than 15 minutes, 15-30 minutes, and more than 30 minutes. A one-way ANOVA was used to determine statistical significance. Tukey's HSD was used post hoc to verify significant differences among observation duration subgroups within the teacher and principal groups. The data from these tests were used to analyze the first and second research questions.

Responses for the survey items related to the Evaluation Feedback construct were analyzed for principals and teachers. The mean for each item was reported to compare the perceptions of principals and teachers within this construct. The mean and standard deviation for all items contributing to the construct were also calculated for each group. A one-way ANOVA was used to determine statistical significance. Tukey's HSD was used post hoc to verify significance among observation duration subgroups. The data from these tests were used to analyze the first and second research questions.

A 3 x 2 factorial design was used to determine statistical significance among the observation duration subgroups within the principal and teacher groups. A univariate general linear model was used for the factorial design. Tukey's HSD was used to test for significance. The analysis was conducted for both survey constructs. Educational role, principal or teacher, served as one grouping in the factorial design. The other grouping was based upon reported duration of observation: less than 15 minutes, 15-30 minutes, and more than 30 minutes. Comparisons were made between the principal and teacher

groups and subgroups using inferential statistics for the two survey constructs to analyze the third research question.

Supporting Research Question 1

Descriptive statistics. Principal responses to the items for both survey constructs were analyzed to address Research Question 1: What are the differences in perceptions among secondary principals regarding the duration of classroom observations conducted by principals and evaluative feedback? H_{01}: Significant differences will not exist in the perceptions of secondary principal subgroups regarding the duration of classroom observations conducted by principals and evaluative feedback.

Survey responses were received from 195 Missouri high school principals representing 35.5% of the population. One hundred eight principals, 55.4%, reported observation durations less than 15 minutes. Seventy-three principals, 37.4%, reported observation durations of 15-30 minutes. Fourteen principals, 7.2%, reported observation durations of more than 30 minutes.

The mean and standard deviation for each item in the two survey constructs were calculated. The potential range for each item was 1-5 with *strongly disagree (*1*), disagree (*2*), neutral (*3*), agree (*4*),* and *strongly agree (*5*).* Table 4 reports the mean and standard deviation for each item. The items are ranked with the highest 50% of the mean values in each construct in bold print.

The mean range of the items in the Duration of Classroom Observations construct was .83 and skewed toward *agree*. Based on the top 50% mean values, principals perceived the greatest agreement that their classroom observations were long enough to assess instructional effectiveness, provide useful feedback, provide quality feedback, and

Table 4

Mean, Standard Deviation, and Range for Principal Responses to Duration of Classroom Observations and Evaluation Feedback Construct Items

Item	*M*	*SD*
Duration of Classroom Observations: Range: 3.28-4.11		
1. The duration of my classroom observations affected the usefulness of the evaluative feedback.	3.52	.91
2. The duration of my classroom observations affected the quality of the evaluative feedback.	3.52	.93
3. The duration of my classroom observations affected the validity of the observations.	3.28	1.02
4. My classroom observations were long enough to assess instructional effectiveness.	**3.94**	.74
5. My classroom observations were long enough to provide useful feedback.	**4.11**	.58
6. My classroom observations were long enough to provide quality feedback.	**4.06**	.58
7. My classroom observations were long enough to accurately reflect instructional practices.	**3.89**	.70
Evaluation Feedback (TEES-P): Range: 3.77-4.14		
1. My evaluation feedback was useful.	**4.01**	.57
2. My evaluation feedback was timely.	**4.04**	.72
3. My evaluation feedback was specific.	**4.13**	.57
4. My evaluation feedback was constructive.	**4.14**	.58
5. My evaluation feedback helped to improve my teachers' instructional effectiveness.	3.90	.57
6. My evaluation feedback represented my teachers' instructional ability.	3.94	.58
7. My evaluation feedback informed specific changes in my teachers' classroom practices.	3.77	.60
8. My evaluation feedback was aligned with the grade level(s) taught by my teachers.	4.00	.67
9. My evaluation feedback was aligned with the subject(s) taught by my teachers.	**4.06**	.67
10. My evaluation feedback was aligned with the school instructional improvement goals.	**4.06**	.69
11. My evaluation feedback was aligned with the school district goals.	**4.08**	.63
12. My evaluation feedback provided information for professional development opportunities for my teachers.	3.81	.76
13. I was satisfied with the feedback I provided during my teachers' evaluations.	3.92	.55

Note. Means in bold represent the top 50% in each construct. n = 191-195. Potential Range = 1-5.

accurately reflect instructional practices. The mean range for the items in the Evaluation Feedback construct was .37 and skewed toward *agree.* Based on the top 50% mean values, principals perceived the greatest agreement that their evaluation feedback was useful, timely, specific, and constructive. Principals also perceived that the evaluation feedback aligned with subjects, school instructional improvement goals, and school district goals.

The mean and standard deviation were calculated for both survey constructs and the two components of the Duration of Classroom Observations construct. Table 5 reports the summative results for all items in the survey constructs (*n* = number of items).

Table 5

Mean, Standard Deviation, and Range of Principal Responses for Construct Groupings

Construct	*n*	*M*	*SD*	Range
Duration of Classroom Observations	7	26.32	3.59	7-35
Component 1	3	10.31	2.61	3-15
Component 2	4	15.99	2.30	4-20
Evaluation Feedback (TEES-P)	13	52.02	4.48	13-65

The Duration of Classroom Observations construct data were non-normally distributed with skewness of -.83 (*SE* = .18). The construct contained seven items with a potential range of 7-35. The mean in the middle of the range was 21. The actual mean was 26.32 (*SD* = 3.59) 95% CI [25.92, 26.87] and was 5.32 above the average mean. The potential range for Component 1 of the construct (Items 1-3) was 3-15. The mean in the middle of the range was 9. The actual mean was 10.31 (*SD* = 2.61) 95% CI [10.00, 10.74] and was 1.31 above the average mean. The potential range for Component 2 of the construct (Items 4-7) was 4-20. The mean in the middle of the range was 12. The

actual mean was 15.99 (*SD* = 2.30) 95% CI [15.70, 16.34] and was 3.99 above the average mean.

The Evaluation Feedback (TEES-P) construct data were near normally distributed with skewness of .05 (*SE* = .18). The construct contained 13 items with a potential range of 13-65. The mean in the middle of the range was 39. The actual mean was 52.02 (*SD* = .48) 95% CI [51.36, 52.69] and was 13.02 above the average mean.

The results for both constructs skewed toward *agree.* The Evaluation Feedback construct had a higher standard deviation and aligned more strongly with *agree* than the Duration of Classroom Observations construct. Component 1 of the Duration of Classroom Observations construct skewed the least away from *neutral.* Component 2 had the lowest standard deviation. The data indicated greater agreement in principal perceptions regarding the adequacy of reported observation durations than effects of observation duration.

Survey responses were evaluated to determine if there were differences in responses based upon reported observation durations. Principals reported the average observation durations in their settings. Possible responses were less than 15 minutes, 15-30 minutes, and more than 30 minutes. Table 6 reports the number of valid responses and means for each group. The mean values reported in Table 6 were used for inferential statistical analysis to explore significant differences in observation duration subgroups.

Inferential statistics. Survey responses for each principal duration of observation subgroup were analyzed for statistical significance. A one-way ANOVA was used to determine statistical significance. Tukey's HSD was used post hoc to verify

significance among observation duration subgroups. Statistically significant differences within the subgroups were not found for any items in the Duration of Classroom

Table 6

Mean of Principal Responses for Construct Groupings by Observation Duration

Construct	Range	< 15 minutes *M*	15-30 minutes *M*	> 30 minutes *M*
Duration of Classroom Observations	7-35	25.95 (*n* = 105)	26.79 (*n* = 72)	26.69 (*n* = 13)
Component 1	3-15	9.98 (*n* = 106)	10.72 (*n* = 72)	10.77 (*n* = 13)
Component 2	4-20	15.95 (*n* = 107)	16.07 (*n* = 72)	15.93 (*n* = 14)
Evaluation Feedback (TEES-P)	13-65	53.31 (*n* = 105)	51.72 (*n* = 67)	51.23 (*n* =13)

Note. *n* = number of valid responses; *M* = mean.

Observations construct. One item in the Evaluation Feedback (TEES-P) construct indicated statistical significance within the subgroups. However, the post hoc test did not verify statistical significance among the subgroups.

Both survey constructs were analyzed for statistical significance. Statistically significant differences were not found within the subgroups for the Duration of Classroom Observations construct, $F = (2, 187) = 1.243$, $p = .291$. Statistically significant differences were not found within the subgroups for the Evaluation Feedback (TEES-P) construct, $F = (2, 182) = .578$, $p = .562$. See Table 6 for the means. Tables 7 and 8 display ANOVA results for the Duration of Classroom Observations and Evaluation Feedback (TEES-P) constructs for principal subgroups.

Table 7

Analysis of Variance for Duration of Classroom Observations Construct Principal Subgroups

	df	*SS*	*MS*	*F*	*p*-value
Between Groups	2	32.01	16.01	1.24	.291*
Within Groups	187	2407.41	12.87		
Total	189	2439.42			

Note. $n = 190$.
*$p > .05$, not significant.

Table 8

Analysis of Variance for Evaluation Feedback (TEES-P) Construct Principal Subgroups

	df	*SS*	*MS*	*F*	*p*-value
Between Groups	2	23.37	11.68	.58	.562*
Within Groups	182	3676.55	20.20		
Total	184	3699.91			

Note. n = 185.
*$p > .05$, not significant.

The following null hypothesis existed for Research Question 1: H_{01}: Significant differences will not exist in the perceptions of secondary principal subgroups regarding the duration of classroom observations conducted by principals and evaluative feedback. The researcher accepted the null hypothesis. Statistically significant differences were not found in the analysis of data. Kruskal-Wallis test results were not significant for either survey construct ($p = .127$ and $p = .244$) and supported acceptance of the null hypothesis.

Principal perceptions regarding the effects of observation duration, adequacy of observation duration, characteristics of feedback, and impact of feedback did not differ significantly according to the reported duration of observations. Responses to Component 1 of the Duration of Classroom Observations construct skewed the least away from *neutral.* Component 2 had the lowest standard deviation. The data indicated greater agreement in principal perceptions regarding the adequacy of reported observation durations than effects of observation duration. Responses to the Evaluation

Feedback Construct (TEES-P) indicated agreement among principals regarding the characteristics of feedback than the impact of their evaluation feedback.

Supporting Research Question 2

Descriptive statistics. Teacher responses to the items for both survey constructs were analyzed to address Research Question 2: What are the differences in perceptions among secondary teachers regarding the duration of classroom observations conducted by principals and evaluative feedback? H_{02}: Significant differences will not exist in the perceptions of secondary teacher subgroups regarding the duration of classroom observations conducted by principals and evaluative feedback.

Survey responses were received from 498 Missouri high school teachers representing 2.5% of the population. Two hundred sixty-five teachers, 53.2%, reported observation durations less than 15 minutes. One hundred sixty-four teachers, 32.9%, reported observation durations of 15-30 minutes. Sixty-nine teachers, 13.9%, reported observation durations of more than 30 minutes.

The mean and standard deviation for each item in the two survey constructs was calculated. The potential range for each item was 1-5 with *strongly disagree (*1*), disagree (*2*), neutral (*3*), agree (*4*),* and *strongly agree (*5*).* Table 9 reports the mean and standard deviation for each item. The items are ranked with the highest 50% of the mean values in each construct in bold print.

The mean range of the items in the Duration of Classroom Observations construct was .39 and skewed away from *neutral*. Based on the three items with the highest mean values, teacher perceptions indicated the greatest agreement with the duration of

Table 9

Mean, Standard Deviation, and Range for Teacher Responses to Duration of Classroom Observations and Evaluation Feedback Construct Items

Item	*M*	*SD*
Duration of Classroom Observations: Range: 3.14-3.53		
1. The duration of my classroom observations affected the usefulness of the evaluative feedback.	**3.47**	.98
2. The duration of my classroom observations affected the quality of the evaluative feedback.	**3.50**	.98
3. The duration of my classroom observations affected the validity of the observations.	**3.53**	1.03
4. My classroom observations were long enough to assess instructional effectiveness.	3.21	1.09
5. My classroom observations were long enough to provide useful feedback.	**3.44**	1.03
6. My classroom observations were long enough to provide quality feedback.	3.38	1.05
7. My classroom observations were long enough to accurately reflect instructional practices.	3.14	1.08
Evaluation Feedback (TEES-T): Range: 2.99-4.05		
1. The evaluation feedback was useful.	3.56	.97
2. The evaluation feedback was timely.	**4.05**	.86
3. The evaluation feedback was specific.	**3.83**	.88
4. The evaluation feedback was constructive.	**3.79**	.90
5. The evaluation feedback helped to improve my instructional effectiveness.	3.36	1.02
6. The evaluation feedback represented my instructional ability.	3.34	1.04
7. The evaluation feedback informed specific changes in my classroom practices.	3.15	.99
8. The evaluation feedback was aligned with the grade level(s) I teach.	**3.82**	.80
9. The evaluation feedback was aligned with the subject(s) that I teach.	**3.69**	.94
10. The evaluation feedback was aligned with the school instructional improvement goals.	**3.90**	.75
11. The evaluation feedback was aligned with the school district goals.	**3.91**	.75
12. The evaluation feedback provided information for professional development opportunities.	2.99	1.05
13. I was satisfied with the feedback I received from my teacher evaluations.	3.59	1.00

Note. Means in bold represent the top 50% in each construct. n = 483-498. Potential Range = 1-5.

classroom observations affecting the usefulness and quality of evaluative feedback and the validity of the observations. The mean range for the items in the Evaluation Feedback construct was 1.06 and skewed toward *agree.* Based on the top 50% mean values, teachers perceived that their evaluation feedback was timely, specific, and constructive. Teachers also perceived that the evaluation feedback aligned with the grade levels and subjects taught, school instructional improvement goals, and school district goals.

The mean and standard deviation were calculated for both survey constructs and the two components of the Duration of Classroom Observations construct. Table 10 reports the summative results for all items in the survey constructs.

Table 10

Mean, Standard Deviation, and Range of Teacher Responses for Construct Groupings

Construct	*n*	*M*	*SD*	Range
Duration of Classroom Observations	7	23.66	4.81	7-35
Component 1	3	10.50	2.73	3-15
Component 2	4	13.16	3.87	4-20
Evaluation Feedback (TEES-T)	13	46.95	9.01	13-65

Note. n = number of survey items in each construct.

The Duration of Classroom Observations construct data was non-normally distributed with skewness of -.46 (*SE* = .11). The construct contained seven items with a potential range of 7-35. The mean in the middle of the range was 21. The actual mean was 23.66 (*SD* = 4.81) 95% CI [23.26, 24.15] and was 2.66 above the average mean. The potential range for Component 1 of the construct (Items 1-3) was 3-15. The mean in the middle of the range was 9. The actual mean was 10.50 (*SD* = 2.73) 95% CI [10.32, 10.82] and was 1.50 above the average mean. The potential range for Component 2 of the construct (Items 4-7) was 4-20. The mean in the middle of the range was 12. The

actual mean was 13.16 (*SD* = 3.87) 95% CI [12.78, 13.50] and was 1.16 above the average mean.

The Evaluation Feedback (TEES-T) construct data were non-normally distributed with skewness of -.66 (*SE* = .11). The construct contained 13 items with a potential range of 13-65. The mean in the middle of the range was 39. The actual mean was 46.95 (*SD* = 9.01) 95% CI [46.09, 47.76] and was 7.95 above the average mean.

The mean for the Evaluation Feedback construct was closer to *agree* than the Duration of Classroom Observations construct with a higher standard deviation. Component 2 of the Duration of Classroom Observations construct had the lowest standard deviation and least deviation from *neutral*. The data indicated greater agreement in teacher perceptions regarding effects of observation duration than the adequacy of reported observation durations.

Survey responses were evaluated to determine if there were differences in responses based upon reported observation durations. Teachers reported the average observation durations in their settings. Possible responses were less than 15 minutes, 15-

Table 11

Mean of Teacher Responses for Construct Groupings by Observation Duration

Construct	Range	< 15 minutes *M*	15-30 minutes *M*	> 30 minutes *M*
Duration of Classroom Observations	7-35	22.00 (*n* = 259)	25.40 (*n* = 161)	25.83 (*n* = 69)
Component 1	3-15	10.22 (*n* = 261)	10.91 (*n* = 163)	10.59 (*n* = 69)
Component 2	4-20	11.79 (*n* =261)	14.49 (*n* = 162)	15.23 (*n* = 69)
Evaluation Feedback (TEES-T)	13-65	44.86 (*n* = 243)	49.73 (*n* = 154)	48.27 (*n* = 60)

Note. *n* = number of valid responses. *M* = mean.

30 minutes, and more than 30 minutes. Table 11 reports the number of valid responses and means for each group.

Inferential statistics. Survey responses for each teacher duration of observation subgroup were analyzed for statistical significance. A one-way ANOVA was used to determine statistical significance. Tukey's HSD was used post hoc to verify significance among observation duration subgroups. Statistically significant differences within the subgroups were found for 5 of the 7 items in the Duration of Classroom Observations construct. These statistically significant differences were found for 1 of the 3 items in Component 1 and for all four items in Component 2. Ten of the 13 items in the Evaluation Feedback (TEES-T) construct indicated statistical significance within the subgroups.

Both survey constructs were analyzed for statistical significance. Statistically significant differences were found within the subgroups for the Duration of Classroom Observations construct, $F = (2, 486) = 38.06$, $p < .001$. Statistically significant differences were also found within the subgroups for Components 1 and 2 of the Duration of Classroom Observations construct, $F = (2, 490) = 3.31$, $p = .037$, and $F = (2, 489) = 41.73$, $p < .001$. Statistically significant differences were found within the subgroups for the Evaluation Feedback (TEES-T) construct, $F = (2, 454) = 15.45$, $p < .001$. See Table 11 for the means. Tables 12-15 display ANOVA results for the Duration of Classroom Observations and Evaluation Feedback (TEES-T) constructs for teacher subgroups.

Post hoc comparisons using Tukey's HSD indicated significant differences in responses between 2 of the 3 possible comparisons in the Duration of Classroom Observations construct. The perceptions of teachers reporting observation durations

Table 12

Analysis of Variance for Duration of Classroom Observations Construct Teacher Subgroups

	df	*SS*	*MS*	*F*	*p*-value
Between Groups	2	1530.30	765.15	38.06	.000*
Within Groups	486	9771.67	20.11		
Total	488	11301.97			

Note. $n = 489$.
*$p \leq .05$.

Table 13

Analysis of Variance for Component 1 of Duration of Classroom Observations Construct Teacher Subgroups

	df	*SS*	*MS*	*F*	*p*-value
Between Groups	2	48.70	24.35	3.31	.037*
Within Groups	490	3608.55	7.36		
Total	492	3657.25			

Note. n = 493.
*$p \leq .05$.

Table 14

Analysis of Variance for Component 2 of Duration of Classroom Observations Construct Teacher Subgroups

	df	*SS*	*MS*	*F*	*p*-value
Between Groups	2	1069.40	534.70	41.73	.000*
Within Groups	489	6265.59	12.81		
Total	491	7334.99			

Note. $n = 492$.
*$p \leq .05$.

Table 15

Analysis of Variance for Evaluation Feedback (TEES-T) Construct Teacher Subgroups

	df	*SS*	*MS*	*F*	*p*-value
Between Groups	2	2357.40	1178.70	15.45	.000*
Within Groups	454	34640.24	76.30		
Total	456	36997.63			

Note. $n = 457$.
*$p \leq .05$.

of less than 15 minutes differed significantly from those reporting durations of 15-30 minutes ($p < .001$; $d = .77$) and those reporting durations of more than 30 minutes ($p <$

.001; d = .83). Significant differences were found in responses between 1 of the 3 comparisons in Component 1 of the construct. The perceptions of teachers reporting observation durations of less than 15 minutes differed significantly from those reporting durations of 15-30 minutes (p = .029; d = .26). Significant differences were found in responses between 2 of the 3 comparisons in Component 2 of the construct. The perceptions of teachers reporting observation durations of less than 15 minutes differed significantly from those reporting durations of 15-30 minutes (p < .001; d = .76) and those reporting duration of more than 30 minutes (p < .001; d = .97). Significant differences were found in responses between 2 of the 3 comparisons in the Evaluation Feedback (TEES-T) construct. The perceptions of teachers reporting observation durations of less than 15 minutes differed significantly from those reporting durations of 15-30 minutes (p < .001; d = .55) and those reporting durations of more than 30 minutes (p = .019; d = .41).

The following null hypothesis existed for Research Question 2: H_{02}: Significant differences will not exist in the perceptions of secondary teacher subgroups regarding the duration of classroom observations conducted by principals and evaluative feedback. The researcher rejected the null hypothesis. Statistically significant differences were found in the analysis of data. Kruskal-Wallis test results were significant for both survey constructs (p < .001) and supported rejection of the null hypothesis.

Teachers experiencing observation durations of less than 15 minutes expressed significantly lower agreement in perceptions regarding the effects of observation duration, adequacy of observation duration, characteristics of feedback, and impact of feedback. The perceptions of teachers experiencing observation durations of less than 15

minutes differed significantly for those experiencing observations of 15-30 minutes and more than 30 minutes for both survey constructs. Large effect sizes were calculated between the less than 15 minutes subgroup and the other two subgroups (d = .77 and d = .83) for the Duration of Classroom Observations construct. Large effects sizes were also calculated between the less than 15 minutes subgroup and the other two subgroups (d = .76 and d = .97) for Component 2 of the construct.

The mean for the Evaluation Feedback (TEES-T) construct was closer to *agree* than the Duration of Classroom Observations construct with a higher standard deviation. Component 2 of the Duration of Classroom Observations construct had the lowest standard deviation and least deviation from *neutral*. The data indicated greater agreement in teacher perceptions regarding effects of observation duration than the adequacy of reported observation durations. Responses to the Evaluation Feedback Construct (TEES-T) indicated greater agreement among teachers regarding the characteristics of evaluation feedback than the impact of their evaluation feedback.

Supporting Research Question 3

Descriptive statistics. Principal and teacher responses to the items for both survey constructs were analyzed to address Research Question 3: What are the differences between the perceptions of secondary principals and teachers regarding the duration of classroom observations conducted by principals and evaluative feedback? H_{03}: Neither principals nor teachers will perceive that the quality of evaluative feedback is affected by the duration of classroom observations conducted by principals. There will be no significant differences between the perceptions of secondary principals and teachers.

Survey responses were received from 195 Missouri high school principals and 498 high school teachers representing 35.5% and 2.5% of the populations respectively. One hundred eight principals, 55.4%, reported observation durations less than 15 minutes. Seventy-three principals, 37.4%, reported observation durations of 15-30 minutes. Fourteen principals, 7.2%, reported observation durations of more than 30 minutes. Two hundred sixty-five teachers, 53.2%, reported observation durations less than 15 minutes. One hundred sixty-four teachers, 32.9%, reported observation durations of 15-30 minutes. Sixty-nine teachers, 13.9%, reported observation durations of more than 30 minutes. The reported observation durations for each population and percentages are reported in Table 16.

Table 16

Reported Observation Durations by Role With Totals and Percentages

Role	< 15 minutes	15-30 minutes	> 30 minutes
Principals (*n*)	108	73	14
Principals (%)	55.4%	37.4%	7.2%
Teachers (*n*)	265	164	69
Teachers (%)	53.2%	32.9%	13.9%
Total (*n*)	373	237	83
Total (%)	53.8%	34.2%	12%

Respondents reported the average frequency of observations conducted or received annually. Respondents selected from ranges of 1-3, 4-6, 7-9, and 10 or more. The range with the highest percentage for principals was 4-6 with a count of 123, 63.1%. The range with the highest percentage for teachers was 1-3 with a count of 236, 47.4%. The range with the highest total percentage was 4-6 with a count of 308, representing 44.4% of all respondents. The reported frequency of observations for each population

and percentages are reported in Table 17. Table 18 shows the cross tabulation of reported observation duration and frequency.

Table 17

Reported Frequency of Observation by Role with Totals and Percentages

Role	1-3	4-6	7-9	10 or more
Principals (*n*)	26	123	40	6
Principals (%)	13.3%	63.1%	20.5%	3.1%
Teachers (*n*)	236	185	65	12
Teachers (%)	47.4%	37.1%	13.1%	2.4%
Total (*n*)	262	308	105	18
Total (%)	37.8%	44.4%	15.2%	2.6%

Table 18

Cross Tabulation of Reported Observation Duration and Observation Frequency

	1-3	4-6	7-9	10 or more	Total
< 15 minutes	112	184	64	13	373
15-30 minutes	84	110	41	2	237
> 30 minutes	66	14	0	3	83
Total	262	308	105	18	693

The reported observation duration with the highest count was less than 15 minutes, $n = 373$, 53.8%. The observation frequency with the highest count was 4-6, $n = 308$, 44.4%. The combination of observation duration and frequency with the highest count was less than 15 minutes and 4-6 annual observations, $n = 184$, 26.6%. The cross tabulation indicated an inverse relationship between observation duration and frequency.

Inferential statistics. A 3 x 2 factorial design was used to determine statistical significance among the observation duration subgroups within the principal and teacher groups. Tukey's HSD was used to test for significance. The analysis was conducted for both survey constructs. Educational role, principal or teacher, served as one grouping in

the factorial design. The other grouping was based upon reported duration of observation: less than 15 minutes, 15-30 minutes, and more than 30 minutes.

A factorial ANOVA was conducted to compare the main effects of educational role and reported observation duration and the interaction effect between the educational role and reported observation duration for the Duration of Classroom Observations construct. The main effect for educational role produced an F ratio of $F(1, 673) = 17.06$, $p < .001$; $d = .63$, indicating a significant difference between principals ($M = 26.32$, $SD = 3.59$) and teachers ($M = 23.66$, $SD = 4.81$). The main effect for reported observation duration produced an F ratio of $F(2, 673) = 17.17$, $p < .001$, indicating significant differences within the groups: less than 15 minutes ($M = 23.14$, $SD = 4.52$), 15-30 minutes ($M = 25.83$, $SD = 4.25$), and more than 30 minutes ($M = 25.96$, $SD = 4.75$). Post hoc comparisons using Tukey's HSD indicated significant differences between 2 of the 3 comparisons. The perceptions of respondents reporting observation durations of less than 15 minutes differed significantly from those reporting durations of 15-30 minutes ($p < .001$; $d = .61$) and those reporting durations of more than 30 minutes ($p < .001$; $d = .61$). The interaction effect between educational role and reported observation duration was significant, $F(2, 673) = 6.23$, $p = .001$.

Table 19

Analysis of Variance Between Educational Role and Reported Observation Duration for the Duration of Classroom Observations Construct

	df	*SS*	*MS*	*F*	*p*-value
Role	1	308.80	308.80	17.06	.000*
Duration	2	621.59	310.79	17.17	.000*
Role*Duration	2	239.80	119.90	6.63	.001*
Error	673	12179.07	18.10		

*$p \leq .05$.

Table 19 reports the analysis of variance between education role and reported observation duration for the Duration of Classroom Observations construct.

A factorial ANOVA was conducted to compare the main effects of educational role and reported observation duration and the interaction effect between the educational role and reported observation duration for Component 1 of the Duration of Classroom Observations construct. The main effect for educational role produced an *F* ratio of $F(1, 678) = .074$, $p = .785$, indicating the difference between principals ($M = 10.31$, $SD = 2.61$) and teachers ($M = 10.50$, $SD = 2.72$) was not significant. The main effect for reported observation duration produced an *F* ratio of $F(2, 678) = 4.53$, $p = .011$, indicating significant difference within the groups: less than 15 minutes ($M = 10.15$, $SD = 2.80$), 15-30 minutes ($M = 10.86$, $SD = 2.44$), and more than 30 minutes ($M = 10.63$, $SD = 2.74$). Post hoc comparisons using Tukey's HSD indicated significant differences between 1 of the 3 comparisons. The perceptions of respondents reporting observation durations of less than 15 minutes differed significantly from those reporting durations of 15-30 minutes ($p \leq .005$; $d = .27$). The interaction effect between educational role and reported observation duration was not significant, $F(2, 678) = .115$, $p = .891$. Table 20 reports the analysis of variance between education role and reported observation duration for Component 1 of the Duration of Classroom Observations construct.

Table 20

Analysis of Variance Between Educational Role and Reported Observation Duration for Component 1 of the Duration of Classroom Observations Construct

	df	*SS*	*MS*	*F*	*p*-value
Role	1	.53	.543	.074	.785**
Duration	2	65.18	32.59	4.53	.011*
Role*Duration	2	1.66	.827	.115	.891**
Error	678	4873.26	7.19		

*$p \leq .05$. **$p \geq .05$, not significant.

A factorial ANOVA was conducted to compare the main effects of educational role and reported observation duration and the interaction effect between the educational role and reported observation duration for Component 2 of the Duration of Classroom Observations construct. The main effect for educational role produced an *F* ratio of $F(1, 679) = 32.47$, $p < .001$; $d = .89$, indicating a significant difference between principals ($M = 15.99$, $SD = 2.30$) and teachers ($M = 13.16$, $SD = 3.87$). The main effect for reported observation duration produced an *F* ratio of $F(2, 679) = 13.72$, $p < .001$, indicating significant difference within the groups: less than 15 minutes ($M = 13.00$, $SD = 4.05$), 15-30 minutes ($M = 14.97$, $SD = 2.91$), and more than 30 minutes ($M = 15.35$, $SD = 2.96$). Post hoc comparisons using Tukey's HSD indicated significant differences between 2 of the 3 comparisons. The perceptions of respondents reporting observation durations of less than 15 minutes differed significantly from those reporting durations of 15-30 minutes ($p < .001$; $d = .56$) and those reporting durations of 30 minutes or more ($p < .001$; $d = .66$). The interaction effect between educational role and reported observation duration was significant, $F(2, 679) = 12.30$, $p < .001$. Table 21 reports the analysis of variance between education role and reported observation duration for Component 2 of the Duration of Classroom Observations construct.

Table 21

Analysis of Variance Between Educational Role and Reported Observation Duration for Component 2 of the Duration of Classroom Observations Construct

	df	*SS*	*MS*	*F*	*p*-value
Role	1	347.90	347.90	32.47	.000*
Duration	2	294.02	147.01	13.72	.000*
Role*Duration	2	263.50	131.75	12.30	.000*
Error	679	7275.94	10.72		

*$p \leq .05$.

A factorial ANOVA was conducted to compare the main effects of educational role and reported observation duration and the interaction effect between the educational role and reported observation duration for the Evaluation Feedback (TEES-P and TEES-T) construct. The main effect for educational role produced an *F* ratio of $F(1, 636) = 19.88$, $p \leq .000$; $d = .71$, indicating a significant difference between principals ($M = 52.02$, $SD = 4.48$) and teachers ($M = 46.95$, $SD = 9.01$). The main effect for reported observation duration produced an *F* ratio of $F(2, 636) = 4.36$, $p = .013$, indicating a significant difference within the groups: less than 15 minutes ($M = 47.11$, $SD = 8.8.$), 15-30 minutes ($M = 50.33$, $SD = 7.47$), and more than 30 minutes ($M = 48.79$, $SD = 6.96$). Post hoc comparisons using Tukey's HSD indicated significant differences between 1 of the 3 comparisons. The perceptions of respondents reporting observation durations of less than 15 minutes differed significantly from those reporting durations of 15-30 minutes ($p < .001$; $d = .39$). The interaction effect between educational role and reported observation duration was significant, $F(2, 636) = 7.54$, $p = .001$. Table 22 reports the analysis of variance between education role and reported observation duration for the Evaluation Feedback (TEES-P and TEES-T) construct.

Table 22

Analysis of Variance Between Educational Role and Reported Observation for the Evaluation Feedback (TEES-P and TEES-T) Construct

	df	*SS*	*MS*	*F*	*p*-value
Role	1	1197.43	1197.43	19.88	.000*
Duration	2	525.25	262.63	4.36	.013*
Role*Duration	2	908.16	454.08	7.54	.001*
Error	636	38316.79	60.25		

*$p \leq .05$.

The following null hypothesis existed for Research Question 3: H_{03}: Neither principals nor teachers will perceive that the quality of evaluative feedback is affected by

the duration of classroom observations conducted by principals. There will be no significant differences between the perceptions of secondary principals and teachers. The researcher rejected the null hypothesis. Statistically significant differences were found in the analysis of data. Kruskal-Wallis test results were significant for both survey constructs ($p < .001$) across categories of role and observation duration. These results supported rejection of the null hypothesis.

The main effect of educational role (principal or teacher) was significant for both survey constructs and Component 2 of the Duration of Classroom Observations construct. The effect size ($d = .89$) was large for Component 2. Educational role did not significantly affect perceptions regarding the effects of observation duration. However, principals perceived significantly greater agreement regarding the adequacy of observation duration, the characteristics of evaluation feedback, and the impact of evaluation feedback. The main effect of reported observation duration was significant for both survey constructs. The perceptions of respondents reporting observation durations of less than 15 minutes indicated significantly lower agreement from those reporting durations of 15-30 minutes regarding the effects of observation duration and the characteristics and impact of evaluation feedback. The perceptions of respondents reporting observation durations of less than 15 minutes also indicated significantly lower agreement than those reporting durations of 15-30 minutes, and 30 or more minutes regarding the adequacy of observation duration. The interaction effect between role and reported observation duration was significant for both survey constructs and Component 2 of the Duration of Classroom Observations construct. It was not significant for Component 1 of the Duration of Classroom Observations construct. The interaction

between role and reported observation duration influenced significant differences in perceptions regarding the adequacy of observation duration, the characteristics of evaluation feedback, and the impact of evaluation feedback. The interaction between role and reported observation duration did not produce significant differences in perceptions regarding the effects of observation duration.

Summary

This chapter presented results of the survey analysis relating to the research questions guiding the study. The mean results for each survey item in both constructs for principals and teachers were ranked and displayed. The mean scores for the Duration of Classroom Observation construct and the Evaluation Feedback (TEES-P and TEES-T) construct were calculated and discussed. Both survey constructs were analyzed for statistically significant responses for both principals and teachers. The main effects of educational role and reported observation duration and the interaction effect between the educational role and reported observation duration were analyzed for statistical significance. Inferential statistical findings were presented and discussed. Table 23 displays significance factors for both survey constructs.

Table 23

Null Hypotheses Determinations for Survey Constructs

	Principals	Teachers	Role	Duration	Role*Duration
Duration of Classroom Observations	Accept	Reject	Reject	Reject	Reject
Component 1	Accept	Reject	Accept	Reject	Accept
Component 2	Accept	Reject	Reject	Reject	Reject
Evaluation Feedback (TEES-P and TEES-T)	Accept	Reject	Reject	Reject	Reject

Acceptance or rejection of the null hypotheses related to the research questions was discussed based upon the analysis of data. Chapter Five presents a summary of the study. The interpretation of the research results, implications of the results, and recommendations for future research are presented.

Chapter Five

Conclusions and Recommendations

Introduction

Educators have increasingly shifted to shorter and more frequent classroom observations. When conducted frequently, mini-observations provide for more sampling of instruction over time and increased opportunity for evaluators to provide feedback to teachers (Marshall, 2013). The usefulness of feedback is affected by factors including relevance, accuracy, timeliness, specificity, and immediacy (Reddy et al., 2016; Reeves, 2009). Feedback has been found to promote professional growth in teachers, leading to improved academic outcomes for students (Campbell, 2013; McEntire, 2010; Shana et al., 2015; Zamary, 2012).

More frequent observation is preferable to less frequent observation (Cohen & Goldhaber, 2016; Donaldson, 2016; Marshall, 2012). However, researchers have not defined the point at which additional observations fail to provide additional feedback for growth. Significant questions remain regarding both the frequency and the duration of classroom observations. Variation has been documented in both practice and recommendations for frequency and duration of observation (Cohen & Goldhaber, 2016). The duration of observation recommended by teacher evaluation experts has varied from as few as 3 minutes to 15 minutes or more (Downey et al., 2004; Marshall, 2013; Marshall & Marshall, 2017). The duration of classroom observations has been cited as an area in need of additional research (Kitendo, 2015).

The purpose of this research study was to describe and compare the perceptions of Missouri secondary principals and teachers regarding the duration of classroom

observations and evaluative feedback. The study provided guidance to evaluators on conducting observations of adequate duration to produce evaluative feedback while minimizing the duration of each classroom observation beyond what is necessary. The study explored both principal and teacher perceptions of the duration necessary for classroom observations to yield high-quality evaluative feedback for professional growth.

This quantitative study utilized a cross-sectional survey of secondary school principals and teachers in the State of Missouri. Respondents were identified as principals or teachers within the survey instrument. Participants were asked to identify the average duration of classroom observations in their settings. Participants also reported the frequency of observation. The remaining survey items used a 5-point Likert scale to measure perceptions about the effects and adequacy of the duration of classroom observations and the characteristics and impact of evaluative feedback. Survey responses were analyzed to describe the perceptions of participants and to determine if significant differences existed depending upon the duration of classroom observations experienced by the participants. Comparisons were also made between the perceptions of principals and teachers about the effects of the duration of classroom observations and evaluative feedback.

Conclusions

Supporting Research Question 1. The researcher accepted the null hypothesis for Research Question 1: H_{01}: Significant differences will not exist in the perceptions of secondary principal subgroups regarding the duration of classroom observations conducted by principals and evaluative feedback. Statistically significant differences were not found in the analysis of survey data from principals.

The results for principal responses to both survey constructs skewed toward *agree*. The Evaluation Feedback (TEES-P) construct responses aligned more strongly with *agree* than the Duration of Classroom Observations construct. The data indicated greater agreement in principal perceptions regarding the adequacy of reported observation durations than effects of observation duration. Responses to the Evaluation Feedback Construct (TEES-P) indicated greater agreement in principal perceptions regarding the characteristics of evaluation feedback than the impact of the feedback.

The means for the items in the Duration of Classroom Observations construct ranged from 3.28 to 4.11 and skewed toward *agree*. Items for Component 2 of the construct had the top 50% mean values in the construct and ranged from 3.89 to 4.11. Principals perceived the greatest agreement to items describing their classroom observations as long enough to assess instructional effectiveness, provide useful feedback, provide quality feedback, and accurately reflect instructional practices. Items for Component 1 of the construct skewed the least away from *neutral* and ranged from 3.28 to 3.52. The items for Component 1 used the question stem "The duration of my classroom observations affected." The items for Component 2 used the question stem "My classroom observations were long enough." The item with the lowest mean, 3.28, was "The duration of my classroom observations affected the validity of the observations." The data indicated greater agreement in principal perceptions regarding the adequacy of reported observation durations than effects of observation duration.

The means for the items in the Evaluation Feedback (TEES-P) construct ranged from 3.77 to 4.14 and skewed toward *agree.* Based on the top 50% mean values, principals perceived the greatest agreement that their evaluation feedback was useful,

timely, specific, and constructive and aligned with subjects, school instructional improvement goals, and school district goals. The item with the lowest mean, 3.77, was "My evaluation feedback informed specific changes in my teachers' classroom practices." The data indicated greater agreement in principal perceptions regarding the characteristics of their evaluation feedback than the impact of the feedback.

The perceptions of principals regarding the effects of observation duration, adequacy of observation duration, characteristics of feedback, and impact of feedback did not differ significantly according to the reported duration of observations. The inferential data analysis was limited by a small sample of principals reporting observation durations of more than 30 minutes. Fourteen principals, 7.2% of respondents in the group, reported observation durations of more than 30 minutes. Seventy-three principals, 37.4%, reported observation durations of 15-30 minutes. One hundred eight principals, 55.4%, reported observation durations less than 15 minutes. Evaluators in schools have increasingly adopted evaluation systems utilizing shorter and more frequent classroom observations (Donaldson, 2016). The distribution of principal responses suggests that secondary schools in Missouri have followed this trend.

The absence of statistically significant differences in perceptions among principals is relevant because of the possibility of bias. Evaluator bias has been noted as a limiting factor affecting the validity and reliability of observations (Bell et al., 2014; Kimball & Milanowski, 2009; Kraft & Gilmour, 2017; Warring, 2015). It is possible that the perceptions of principals are influenced by a positive bias toward their ability to assess instruction and deliver feedback. Across the observation duration subgroups, principal perceptions indicated agreement that observations were long enough to assess

instruction and generate useful feedback. While still positive, principals indicated lower levels of agreement regarding the impact of their feedback on changes in instructional practices and the professional development of teachers. Again, this could reflect a bias in principal perceptions that their feedback should produce changes in practice, but teachers are unable or unwilling to act on the feedback. It could also be indicative of the feedback lacking the characteristics needed for teachers to implement changes in practices. The perceptions of principals expressed in this study reinforce the need for evaluator training to include the use of coaching feedback to promote teacher growth (Kimball & Milanowski, 2009; Kraft & Gilmour, 2016; Patrick & Mantzicopoulos, 2016).

Supporting Research Question 2. The researcher rejected the null hypothesis for Research Question 2: H_{02}: Significant differences will not exist in the perceptions of secondary teacher subgroups regarding the duration of classroom observations conducted by principals and evaluative feedback. Statistically significant differences were found in the analysis of survey data from teachers.

The results for teacher responses to the Duration of Classroom Observations construct skewed away from *neutral*. The data indicated greater agreement in teacher perceptions regarding the effects of observation duration than the adequacy of reported observation durations. Responses to the Evaluation Feedback Construct (TEES-T) skewed toward *agree* and indicated greater agreement in teacher perceptions regarding the characteristics of evaluation feedback than the impact of the feedback.

The means for the items in the Duration of Classroom Observations construct ranged from 3.14-3.53 and skewed away from *neutral*. The items for Component 1 of the construct had the highest mean values in the construct and ranged from 3.47 to 3.53.

Teachers perceived the greatest agreement to items describing the duration of classroom observations affecting the usefulness and quality of evaluative feedback and the validity of the observations. Items for Component 2 of the construct skewed the least away from *neutral* and ranged from 3.14 to 3.44. The items for Component 1 used the question stem "The duration of my classroom observations affected." The items for Component 2 used the question stem "My classroom observations were long enough." The item with the lowest mean, 3.14, was "My classroom observations were long enough to accurately reflect instructional practices." The data indicated greater agreement in teacher perceptions regarding the effects of observation duration than the adequacy of reported observation durations.

The means for the items in the Evaluation Feedback (TEES-T) construct ranged from 2.99 to 4.05 and skewed toward *agree.* Based on the top 50% mean values, teachers perceived the greatest agreement that their evaluation feedback was timely, specific, and constructive and aligned with the grade levels and subjects taught, school instructional improvement goals, and school district goals. The item with the lowest mean, 2.99, was "The evaluation feedback provided information for professional development opportunities." The data indicated greater agreement in teacher perceptions regarding the characteristics of the evaluation feedback they received than the impact of the feedback.

The data analysis was limited by teacher responses representing only 2.5% of the secondary teacher population in Missouri. However, the sample was large enough for analysis and to generate significant results. The distribution of teacher responses among the duration of observation subgroups suggests that secondary schools in Missouri have

followed the trend toward shorter and more frequent classroom observations. Sixty-nine teachers, 13.9% of respondents in the group, reported observation durations of more than 30 minutes. One hundred sixty-four teachers, 32.9%, reported observation durations of 15-30 minutes. Two hundred sixty-five teachers, 53.2%, reported observation durations less than 15 minutes. As with principals, teacher perceptions could also be limited by bias associated with positive or negative evaluation experiences. Perceptions could also be influenced by the mindset, efficacy, and ability of teachers to engage in self-reflection (Costa & Garmston, 2015; Dweck, 2016; Kimball & Milanowski, 2009).

Teacher perceptions regarding the adequacy of observation duration, characteristics of feedback, and impact of feedback differed significantly depending upon the reported duration of observations. The perceptions of teachers reporting observation durations less than 15 minutes differed significantly from those reporting durations of 15-30 minutes and more than 30 minutes for both survey constructs. Significant differences were not found between the perceptions of teachers reporting observation durations of 15-30 minutes and those reporting 30 or more minutes. Teachers receiving observations of less than 15 minutes expressed significantly lower agreement regarding the adequacy of observation duration, characteristics of feedback, and impact of feedback compared to those receiving longer observations. The duration of classroom observations is a significant factor in teacher perceptions of classroom observations and evaluation feedback. Evaluators should consider observation duration as a factor influencing teacher perceptions regarding the validity of the observation. Evaluators should also be aware that teachers might have positive perceptions about the characteristics of evaluative feedback while not perceiving that the feedback informs changes in practice. Feedback

should be paired with opportunities to practice instructional strategies in support of changes in practice (Hattie, 2012; McNulty, 2011; Reeves, 2009).

Supporting Research Question 3. The researcher rejected the null hypothesis for Research Question 3: H_{03}: Neither principals nor teachers will perceive that the quality of evaluative feedback is affected by the duration of classroom observations conducted by principals. There will be no significant differences between the perceptions of secondary principals and teachers. Statistically significant differences were found in the analysis of main effects and interaction effects for educational role (principal or teacher) and duration of observation.

Educational role was not found to significantly affect perceptions regarding the effects of observation duration as measured in Component 1 of the Duration of Classroom Observations construct. Principals and teachers did not differ significantly in their perceptions regarding the effects of the duration of their classroom observations. Additionally, the interaction effect between role and reported observation duration was not significant for Component 1 of the Duration of Classroom Observations construct. The interaction between role and reported observation duration did not produce significant differences in perceptions regarding the effects of observation duration.

The main effect of educational role was significant for the Duration of Classroom Observations construct, Component 2 of the construct, and the Evaluation Feedback (TEES-P and TEES-T) construct. Principals perceived significantly greater agreement compared to teachers regarding the adequacy of observation duration, the characteristics of evaluation feedback, and the impact of evaluation feedback.

The main effect of reported observation duration was also significant for both survey constructs. The perceptions of respondents reporting observation durations of less than 15 minutes indicated significantly lower agreement compared to those reporting durations of 15-30 minutes regarding the effects of observation duration and the characteristics and impact of evaluation feedback. The perceptions of respondents reporting observation durations of less than 15 minutes also indicated significantly lower agreement compared to those reporting durations of 15-30 minutes, and 30 or more minutes regarding the adequacy of observation duration. The interaction effect between role and reported observation duration was also significant for both survey constructs and Component 2 of the Duration of Classroom Observations construct. The interaction between role and reported observation duration influenced significant differences in perceptions regarding the adequacy of observation duration, the characteristics of evaluation feedback, and the impact of evaluation feedback.

Principals and teachers reported the average frequency of observation in addition to the average duration of observation. The effects of the frequency of observation and interactions between duration and frequency were outside of the scope of this study. However, the reported frequency of observation aided in describing and comparing evaluation practices. The range with the highest percentage reported by principals was 4-6 observations per teacher annually with a count of 123, 63.1%. The range with the highest percentage reported by teachers was 1-3 observations received annually with a count of 236, 47.4%. The range with the highest total percentage was 4-6 with a count of 308, representing 44.4% of all respondents. The percentages for reported observation durations were similar for both principals and teachers. The reported frequencies of

observation were higher for principals than teachers. The reported frequencies of observation could be accurate. However, reported frequency of observation could also have been influenced by respondent bias. Principals could have perceived that they observed more frequently than they did in practice. Teachers could have perceived that they were observed less frequently than they were in practice. It is also possible that the frequency of observation for teachers varied according to factors such as tenure or years of experience while the duration of observation did not. Researchers have described variation in observation duration and frequency both within and among schools (Cohen & Goldhaber, 2016). The reported durations and frequencies of observation indicate these types of variations exist in Missouri secondary schools. The reported durations and frequencies of observation also support the trend toward shorter and more frequent observations in Missouri secondary schools. An inverse relationship between frequency and duration was noted with frequency generally increasing as duration decreased.

Principals and teachers did not differ significantly in their perceptions that the duration of classroom observation did affect the usefulness and quality of evaluative feedback and the validity of observations. The two groups did differ significantly in their perceptions that the duration of their observations was adequate to provide useful feedback and to accurately assess instructional practices. The perceptions of teachers reporting observation durations less than 15 minutes differed significantly from those reporting longer observation durations. Significant differences were not present in the perceptions of principals. Bias could have been a factor influencing the perceptions of both principals and teachers. However, the results of this study support the concept of the duration of classroom observations affecting the validity of the observations. The

results also support the use of formative feedback and coaching to support professional development and changes in instructional practices (Mette et al., 2017).

Recommendations

Evaluators should consider 15 minutes as a minimum duration and 30 minutes as a maximum duration for classroom observations. Evaluators should be aware that teacher confidence in the validity of observations is likely to increase with longer observations. However, observations longer than 30 minutes do not appear to be necessary according to the results of this study paired with the Measures of Effective Teaching study showing high correlations between 30-minute lesson segments (Joe et al., 2014). Evaluators using observation durations of 15 minutes or less should work locally with teachers to understand their perceptions and adjust observation durations accordingly. Specific recommendations regarding frequency of observation are outside the scope of this study. However, increased frequency of observation is associated with positive outcomes (Kitendo, 2015; McEntire, 2010; Shana et al., 2015; Zamary, 2012). Evaluators should pair 15 to 30-minute timeframes with observations conducted as frequently as is practical. Evaluators should use rubrics to assess a few specific aspects of teaching and learning in each classroom observation (Joe et al., 2014). A feedback conversation between the evaluator and teacher should follow each observation (Coggshall et al., 2012; Downey et al., 2004).

School districts should implement evaluator training procedures if evaluator training is not already in place. Evaluator training should emphasize the use of feedback to inform teacher professional growth and changes in practice. Additionally, evaluator training should emphasize sources of evaluator bias and guidance on minimizing bias

(Murphy & Beretvas, 2015). School districts should develop procedures for ensuring that observation protocols are implemented with fidelity after evaluator training has occurred (Bell et al., 2014; Park et al., 2014). School districts should pair evaluator training with training to assist teachers in understanding evaluation systems. The use of evaluation feedback to promote self-reflection and identify professional growth opportunities should be emphasized in teacher training. Formative feedback, along with other measures of teacher effectiveness, should be used to inform the development of teacher professional growth plans with specific goals for growth (Costa & Garmston, 2015; Mette et al., 2017; Walkowiak, 2016). Teacher growth plans should be supported by coaching opportunities for teachers to practice and improve specific instructional strategies (Hattie, 2012; McNulty, 2011; Reeves, 2009). Finally, school districts should use instruments, such as the Teacher Evaluation Experience Scale, to periodically gather input on teacher perceptions of evaluation systems (Reddy et al., 2016). This input could be used to support teacher confidence in evaluation systems by addressing areas perceived to need improvement.

Future Research Topics

Future research should be conducted to better understand how the duration of classroom observations affects principal and teacher perceptions of evaluative feedback. The study should be replicated with principals and teachers at middle and elementary levels. Methodologies researching smaller ranges of observation duration, such as 10-15 minutes and increments from 15-30 minutes, would refine the results of this study. Methodologies capturing the actual duration and frequencies of observation, rather than reported ranges, would contribute to the objectivity of results. Mechanisms for sampling

a greater percentage of the teacher population would enhance the generalizability of results. The addition of qualitative methods to explore factors underlying principal and teacher perceptions would also expand on the results of this study.

The study should be replicated to research the effects of observation frequency on principal and teacher perceptions of evaluative feedback. The Duration of Classroom Observations construct items could be amended to measure perceptions about the frequency of classroom observations. The researcher recommends that validity and reliability testing be repeated before amended survey items are used. Duration of Classroom Observations and Frequency of Classroom Observations constructs could be paired to research the effects of and interactions between duration and frequency.

Summary

This study indicated the transition to more frequent, shorter classroom observations has occurred in Missouri secondary schools. The perceptions of principals and teachers differed significantly regarding the adequacy of observation durations in accurately assessing instructional practices and providing useful feedback. Principals indicated significantly greater agreement regarding aspects of observation duration and evaluative feedback compared to the perceptions of teachers. Teachers reporting observation durations less than 15 minutes indicated significantly lower agreement compared to teachers reporting longer observation durations. There were no significant differences in the perceptions of principals depending upon reported observation duration. The results of this study indicate the duration of classroom observation affects teacher perceptions regarding the validity of classroom observations and impact of

evaluative feedback. The results also suggest that positive bias could influence the perceptions of principals.

This chapter summarized the purpose, methodology, and data analysis of the study. Differences in perceptions based upon the duration of classroom observations and educational roles of principals and teachers were discussed. Conclusions and recommendations for teacher evaluation practices were offered based upon the review of literature and analysis of data. Recommendations for future research to expand on the results of this study were offered.

References

Aguilar, C., & Richerme, L. (2014). What is everyone saying about teacher evaluation? Framing the intended and inadvertent causes and consequences of Race to the Top. *Arts Education Policy Review*, *115*(4), 110-120.

Akhavan, N., & Tracz, S. (2016). The effects of coaching on teacher efficacy, academic optimism and student achievement: The consideration of a continued professional development option for teachers. *Journal of Education and Human Development, 5*(3), 38-53.

Anderman, E., Gimbert, B., O'Connell, A., & Riegel, L. (2015). Approaches to academic growth assessment. *British Journal of Educational Psychology*, *85*(2), 138-153.

Archer, J., Cantrell, S., Holtzman, S., Joe, J., Tocci, C., &, Wood, J. (2016). *Better feedback for better teaching: A practical guide to improving classroom observations.* San Francisco, CA: Jossey-Bass.

Bailey, M. (2016). Invasion of the body snatchers. *Curriculum & Teaching Dialogue, 18*(1/2), 139-154.

Ballou, D., & Springer, M. (2015). Using student test scores to measure teacher performance: Some problems in the design and implementation of evaluation systems. *Educational Researcher, 44*(2), 77-86.

Beets, P. (2012). Strengthening morality and ethics in educational assessment through ubuntu in South Africa. *Educational Philosophy & Theory, 44*(S2), 68-83.

Bell, C., Qi, Y., Croft, A., Leusner, D., McCaffrey, D., Gitomer, D., & Pianta, R. (2014). Improving observational score quality: Challenges in observer thinking. In T. Kane, K. Kerr, & R. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the Measures of Effective Teaching Project* (pp. 50-97). San Francisco, CA: Jossey-Bass.

Bergin, C. (2015). Using student achievement data to evaluate teachers. Columbia, MO: Network for Educator Effectiveness.

Bill & Melinda Gates Foundation. (2010). *Learning about teaching: Initial findings from the measures of effective teaching project.* Seattle, WA: Author.

Bolman, L., & Deal, T. (2008). *Reframing organizations: Artistry, choice, and leadership* (4th ed.). San Francisco, CA: Jossey-Bass.

Bolyard, C. (2015). Test-based teacher evaluations: Accountability vs. responsibility. *Philosophical Studies in Education, 46,* 73-82.

Brandt, R. (1995). Teacher evaluation for incentive pay and career ladder programs. In D. Duke (Ed.), *Teacher evaluation policy: From accountability to professional development* (pp. 13-34). Albany, NY: State University of New York Press.

Campbell, T. F. (2013). *Teacher supervision and evaluation: A case study of administrators' and teachers' perceptions of mini observations* (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global. (3556916)

Castellano, K., & McCaffrey, D. (2017). The accuracy of aggregate student growth percentiles as indicators of educator performance. *Educational Measurement: Issues & Practice, 36*(1), 14-27.

Chin, M., & Goldhaber, D. (2015). Impacts of multidimensionality and error: Simulating explanations for weak correlations between measures of teacher quality. *Proceedings of the Society for Research on Educational Effectiveness Spring 2015 Conference.* Evanston, IL: Society for Research on Educational Effectiveness.

Coggshall, J., Rasmussen, C., Colton, A., Milton, J., & Jacques, C. (2012). *Generating teaching effectiveness: The role of job-embedded professional learning in teacher evaluation.* Washington, DC: National Comprehensive Center for Teacher Quality.

Cohen, J., & Goldhaber, D. (2016). Building a more complete understanding of teacher evaluation using classroom observations. *Educational Researcher, 45*(6), 378-387.

Colvin, G., Flannery, K. B., Sugai, G., & Monegan, J. (2009). Using observational data to provide performance feedback to teachers: A high school case study. *Preventing School Failure, 53*(2), 95-104.

Costa, A., & Garmston, R. (2015). Check your gauges: Calibrating conversations assist teachers in fine-tuning instruction. *Journal of Staff Development: The Learning Forward Journal, 36*(1), 44-47.

Costa, A., & Garmston, R. with Hayes, C., & Ellison, J. (2016). *Cognitive coaching: Developing self-directed leaders and learners* (3rd ed.). Lanham, MD: Rowman & Littlefield.

Croft, S., Roberts, M., & Stenhouse, V. (2016). The perfect storm of education reform: High-stakes testing and teacher evaluation. *Social Justice, 42*(1), 70-92.

Cusick, P. A. (2014). The logic of the U.S. educational system and teaching. *Theory Into Practice, 53*(3), 176-182.

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

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

Darrow, A-A. (2016). The Every Student Succeeds Act (ESSA): What it means for students with disabilities and music educators. *General Music Today, 30*(1), 41-44.

Donaldson, M. (2016). Teacher evaluation reform: Focus, feedback, and fear. *Educational Leadership, 73*(8), 72-76.

Downey, C., Steffy, B., English, F., Frase, L., & Poston, W., Jr. (2004). *The three-minute classroom walk-through: Changing school supervisory practice one teacher at a time.* Thousand Oaks, CA: Corwin Press.

Duffy, M., Giordano, V., Farrell, J., Paneque, O., & Crump, G. (2008). No Child Left Behind: Values and research issues in high-stakes assessments. *Counseling & Values, 53*(1), 53-66.

Dweck, C. (2016). *Mindset: The new psychology of success.* New York, NY: Ballantine Books.

Ellis, C. R. (2007). No Child Left Behind: A critical analysis. *Curriculum & Teaching Dialogue, 9*(1/2), 221-233.

Feeney, E. J. (2007). Quality feedback: The essential ingredient for teacher success. *Clearing House, 80*(4), 191-198.

Ferguson, R. (2012). Can student surveys measure teaching quality? *Phi Delta Kappan, 94*(3), 24-28.

Ferguson, R., & Danielson, C. (2014). How framework for teaching and Tripod 7Cs evidence distinguish key components of effective teaching. In T. Kane, K. Kerr, & R. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the measures of effective teaching project* (pp. 98-143). San Francisco, CA: Jossey-Bass.

Garmston, R., & Wellman, B. (2009). *The adaptive school: A sourcebook for developing collaborative groups* (2nd ed.). Norwood, MA: Christopher-Gordon.

Glickman, C., Gordon, S., & Ross-Gordon, J. (2010). *Supervision and instructional leadership: A developmental approach* (8th ed.). Boston, MA: Allyn & Bacon.

Goe, L. (2013). Can teacher evaluation improve teaching? *Principal Leadership, 13*(7), 24–28.

Gottlieb, D. (2013). Eisner's evaluation in the age of Race to the Top. *Curriculum & Teaching Dialogue*, *15*(1/2), 11-25.

Hargreaves, A., & Shirley, D. (2009). *The fourth way: The inspiring future for educational change.* Thousand Oaks, CA: Corwin.

Hattie, J. (2012). *Visible learning for teachers: Maximizing impact on learning.* New York, NY: Routledge.

Hattie, J., & Clinton, J. (2011). School leaders as evaluators. In E. Allison, J. Clinton, J. Hattie, C. Kamm, C. Lassiter, B. McNulty...S. White, *Activate: A leader's guide to people, practices, and processes* (pp. 93-118). Englewood, CO: Lead + Learn Press.

Hof, M. (2012). Questionnaire evaluation with factor analysis and Cronbach's alpha: An example. Retrieved from http://www.let.rug.nl/nerbonne/teach/rema-stats-meth-seminar/student-papers/MHof-QuestionnaireEvaluation-2012-Cronbach-FactAnalysis.pdf

Joe, J., McClean, C., & Holtzman, S. (2014). Scoring design decisions: Reliability and the length and focus of classroom observations. In T. Kane, K. Kerr, & R. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the measures of effective teaching project* (pp. 415-443). San Francisco, CA: Jossey-Bass.

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

Kimball, S., & Milanowski, A. (2009). Examining teacher evaluation validity and leadership decision making within a standards-based evaluation system. *Educational Administration Quarterly, 45*(1), 34-70.

Kitendo, P. (2015). *Effect of frequent administrators' walkthrough observations on teachers' classroom instructional practices and students' performance in public schools* (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global. (3718218)

Kostogriz, A., & Doecke, B. (2013). The ethical practice of teaching literacy: Accountability or responsibility? *Australian Journal of Language & Literacy, 36*(2), 90-98.

Kraft, M. & Gilmour, A. (2016). Can principals promote teacher development as evaluators? A case study of principals' views and experiences. *Educational Administration Quarterly, 52*(5), 711-753.

Kraft, M. & Gilmour, A. (2017). Revisiting the widget effect: Teacher evaluation reforms and the distribution of teacher effectiveness. *Educational Researcher 46(*5), 234-249.

Lassiter, C. (2011). Leadership for a high performance culture. In E. Allison, J. Clinton, J. Hattie, C. Kamm, C. Lassiter, B. McNulty...S. White (Eds.), *Activate: A leader's guide to people, practices, and processes* (pp. 57-90). Englewood, CO: Lead + Learn Press.

LaVenia, M., Cohen-Vogel, L., & Lang, L. (2015). The Common Core State Standards initiative: An event history analysis of state adoption. *American Journal of Education*, *121*(2), 145-182.

LeFevre, D., & Robinson, V. (2015). The interpersonal challenges of instructional leadership: Principals' effectiveness in conversations about performance issues. *Educational Administration Quarterly, 51*(1), 58- 95.

Marsh, J. A., Bush-Mecenas, S., & Hough, H. (2017). Learning from early adopters in the new accountability era: Insights from California's CORE waiver districts. *Educational Administration Quarterly*, *53*(3), 327-364.

Marshall, K. (2012). Let's cancel the dog-and-pony show. *Phi Delta Kappan*, *94*(3), 19-23.

Marshall, K. (2013). *Rethinking teacher supervision and evaluation: How to work smart, build collaboration, and close the achievement gap* (2nd ed.). San Francisco, CA: Jossey-Bass.

Marshall, K., & Marshall, D. (2017). Mini-observations: A keystone habit. *School Administrator, 74*(110), 26-29.

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

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

Maslow, A. (1954). *Motivation and personality.* New York, NY: Harper & Row.

McDowall, A., Freeman, K., & Marshall, S. (2014). Is FeedForward the way forward? A comparison of the effects of FeedForward coaching and feedback. *International Coaching Psychology Review, 9*(2), 135-146.

McEntire, L. (2010). *The use of classroom walk-through observations as a strategy for improving teaching and learning: Teacher perspective* (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global. (3389734)

McGregor, D. (1960). *The human side of enterprise.* New York, NY: McGraw-Hill.

McNulty, B. (2011). Leaders developing learning systems. In E. Allison, J. Clinton, J. Hattie, C. Kamm, C. Lassiter, B. McNulty…S. White (Eds.), *Activate: A leader's guide to people, practices, and processes* (pp. 171-199). Englewood, CO: Lead + Learn Press.

Mette, I., Anderson, J., Nieuwenhuizen, L., Range, B., Hvidston, D., & Doty, J. (2017). The wicked problem of the intersection between supervision and evaluation. *International Electronic Journal of Elementary Education*, *9*(3), 709-724.

Missouri Department of Elementary and Secondary Education. (2018). *2016-2017 Statistics of Missouri public schools* [Data file]. Retrieved from https://mcds.dese.mo.gov/quickfacts/District%20and%20School%20Information/Missouri%20School%20Statistics.pdf

Murphy, D., & Beretvas, S. (2015). A comparison of teacher effectiveness measures calculated using three multilevel models for raters effects. *Applied Measurement in Education*, *28*(3), 219-236.

National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform.* Washington, DC: Author.

Neumerski, C. (2013). Rethinking instructional leadership, a review: What do we know about principal, teacher, and coach instructional leadership, and where should we go from here? *Educational Administration Quarterly*, *49*(2), 310-347.

The New Teacher Project. (2010). *How federal education policy can reverse the widget effect: Transforming ESEA Title II to improve teacher effectiveness and student outcomes.* Brooklyn, NY: Author.

No Child Left Behind Act of 2001, P.L. 107-110, 20 U.S.C. § 6319 (2002).

Park, Y., Chin, J., & Holtzman, S. (2014). Evaluating efforts to minimize rater bias in scoring classroom observations. In T. Kane, K. Kerr, & R. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the measures of effective teaching project* (pp. 383-414). San Francisco, CA: Jossey-Bass.

Patrick, H., & Mantzicopoulos, P. (2016). Is effective teaching stable? *Journal of Experimental Education, 84*(1), 23-47.

Polikoff, M. (2015). The stability of observational and student survey measures of teaching effectiveness. *American Journal of Education, 121*(2), 183-212.

Popham, W. J. (2013). Teeter-totters have two ends. *Measurement, 11*(4), 189-191.

Raudenbush, S., & Marshall, J. (2014). To what extent do student perceptions of classroom quality predict teacher value added? In T. Kane, K. Kerr, & R. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the measures of effective teaching project* (pp. 170-202). San Francisco, CA: Jossey-Bass.

Reddy, L., Dudek, C., Kettler, R., Kurz, A., & Peters, S. (2016). Measuring educators' attitudes and beliefs about evaluation: Construct validity and reliability of the teacher evaluation experience scale. *Educational Assessment, 21*(2), 120-134.

Reeves, D. (2009). *Leading change in your school: How to conquer myths, build commitment, and get results.* Alexandria, VA: Association for Supervision and Curriculum Development.

Reinke, W., Stormont, M., Herman, K., & Newcomer, L. (2014). Using coaching to support teacher implementation of classroom-based interventions. *Journal of Behavioral Education, 23*(1), 150-167.

Romano, V. A., Jr. (2014). Secondary teachers' and their supervisors' perceptions of current and desired observation practices. *Global Education Review, 1*(3), 135-146.

Rovinelli, R., & Hambleton, R. (1977). On the use of content specialists in the assessment of criterion-referenced test item validity. *Dutch Journal of Educational Research*, 2, 49-60.

Rutkowski, D., & Wild, J. (2015). Stakes matter: Student motivation and the validity of student assessments for teacher evaluation. *Educational Assessment, 20*(3), 165-179.

Schweig, J. (2014). Quantifying error in survey measures of school and classroom environments. *Applied Measurement in Education, 27*(2), 133-157.

Senge, P. (2006). *The fifth discipline: The art and practice of the learning organization.* New York, NY: Crown Business.

Sergiovanni, T., & Starratt, R. (1993). *Supervision: A redefinition* (5th ed.). New York, NY: McGraw-Hill.

Shana, S., Glassett, K., & Copas, A. (2015). The impact of teacher observations with coordinated professional development on student performance: A 27-state program evaluation. *Journal of College Teaching & Learning, 12*(1), 55-64.

Shapiro, J., & Gross, S. (2013). *Ethical educational leadership in turbulent times: (Re)solving moral dilemmas* (2nd ed.). New York, NY: Routledge.

Snow, S. (2014*). Supervisory perceptions of teacher supervision and effects on student achievement in Missouri* (Doctoral dissertation). Retrieved from http://libguides.sbuniv.edu/c.php?g=154859&p=1087326

Stone, D., & Heen, S. (2014). *Thanks for the feedback: The science and art of receiving feedback well.* New York, NY: Penguin Books.

Tanner, D. (2013). Race to the Top and leave the children behind. *Journal of Curriculum Studies, 45*(1), 4-15.

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

U.S. Department of Education, National Center for Education Statistics, Schools and Staffing Survey (SASS). (2016). *Public school principal data file, 2011-12*. Retrieved from https://nces.ed.gov/datalab/sass/index.aspx

Van der Lans, R., Van de Grift, W., & van Veen, K. (2015). Developing a teacher evaluation instrument to provide formative feedback using student ratings of teaching acts. *Educational Measurement: Issues & Practice, 34*(3), 18-27.

Van der Lans, R., Van de Grift, W., van Veen, K., & Fokkens-Bruinsma, M. (2016). Once is not enough: Establishing reliability criteria for feedback and evaluation decisions based on classroom observations. *Studies in Educational Evaluation, 50,* 88-95.

Walkowiak, T. (2016). Five essential practices for communication: The work of instructional coaches. *Clearing House*, *89*(1), 14-17.

Warring, D. (2015). Teacher evaluations: Use or misuse? *Universal Journal of Educational Research, 3*(10), 703-709.

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness* (2nd ed.). Brooklyn, NY: New Teacher Project. Retrieved from http://tntp.org/assets/documents/TheWidgetEffect_2nd_ed.pdf

Wise, A., Darling-Hammond, L., McLaughlin, M., & Bernstein, H. (1984). *Teacher evaluation: A study of effective practices.* Santa Monica, CA: Rand.

Zamary, J. (2012). *Mini-observations case study: Assessing and providing feedback that can lead to changes in instructional practice* (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global. (3515604)

Appendix A

Draft of Survey Instrument

The purpose of this survey is to explore the perceptions of principals and teachers about the duration of classroom observations conducted by principals and evaluative feedback. Participants should give the response that best reflects their own situation and experiences over the past year.

Demographics and Background Information

1. What is your professional role?
    a. Teacher
    b. Principal
2. What is the average duration of each classroom observation you conduct or receive for evaluative purposes?
    a. Less than 15 minutes
    b. 15 – 30 minutes
    c. More than 30 minutes
3. For principals, what is the average number of times each teacher in your school is observed annually for evaluative purposes?

    For teachers, what is the average number of times that you are observed annually for evaluative purposes?
    a. 1 – 3
    b. 4 – 6
    c. 7 – 9
    d. 10 or more

Participants will respond to Items 4 – 26 on a Likert scale:

1 – strongly disagree

2 – disagree

3 – neutral

4 – agree

5 – strongly agree

Duration of Classroom Observations

4. The duration of my classroom observations affected the usefulness of evaluative feedback.
5. The duration of my classroom observations affected the quality of the evaluative feedback.
6. The duration of my classroom observations affected the validity of the observations.
7. My classroom observations were long enough to assess instructional effectiveness.
8. My classroom observations were long enough to provide useful feedback.
9. My classroom observations were long enough to provide quality feedback.
10. My classroom observations were long enough to accurately reflect instructional practices.
11. Part of an observed lesson represented the quality of the whole lesson.
12. I would have preferred for my classroom observations to be longer in duration.
13. I would have preferred shorter observations over longer observations.

Evaluation Feedback (TEES-T)

14. The evaluation feedback was useful.
15. The evaluation feedback was timely.
16. The evaluation feedback was specific.
17. The evaluation feedback was constructive.
18. The evaluation feedback helped to improve my instructional effectiveness.
19. The evaluation feedback represented my instructional ability.
20. The evaluation feedback informed specific changes in my classroom practices.
21. The evaluation feedback was aligned with grade level(s) I teach.
22. The evaluation feedback was aligned with the subject(s) that I teach.
23. The evaluation feedback was aligned with the school instructional improvement goals.
24. The evaluation feedback was aligned with the school district goals.
25. The evaluation feedback provided information for professional development opportunities.
26. I was satisfied with the feedback I received from my teacher evaluations.

Evaluation Feedback (TEES-P)

14. My evaluation feedback was useful.
15. My evaluation feedback was timely.
16. My evaluation feedback was specific.
17. My evaluation feedback was constructive.

18. My evaluation feedback helped to improve my teachers' instructional effectiveness.
19. My evaluation feedback represented my teachers' instructional ability.
20. My evaluation feedback informed specific changes in my teachers' classroom practices.
21. My evaluation feedback was aligned with the grade level(s) taught by my teachers.
22. My evaluation feedback was aligned with the subject(s) taught by my teachers.
23. My evaluation feedback was aligned with the school instructional improvement goals.
24. My evaluation feedback was aligned with the school district goals.
25. My evaluation feedback provided information for professional development opportunities for my teachers.
26. I was satisfied with the feedback I provided during my teachers' evaluations.

Appendix B

Permission to Use TEES-T and TEES-P

RE: TEES-T

Linda Reddy <lreddy@gsapp.rutgers.edu>

Fri 2/9/2018 4:52 PM

To:David Pyle <s549843@sbuniv.edu>;

Cc:Chris Dudek <cdudek@scarletmail.rutgers.edu>; Ryan Kettler <r.j.kettler@rutgers.edu>; Alexander Kurz <Alexander.Kurz@asu.edu>;

1 attachments (101 KB)

TEES Scale & Items_Rugers_9.10.17.pdf;

Hi David,

Thanks for your message. We would be happy to give you permission to use our scales. I am including my colleagues on this message as I am away for the next 8 days for work. Please find attached the TEES Teacher and School Administrator Forms.

Best,

Linda Reddy

Appendix C

TEES-T and TEES-P Evaluation Feedback Construct

Teacher Evaluation Experiences Survey – Teacher Form (TEES-T) Reddy, Dudek, Kettler, Kurz, & Peters © 2015 Rutgers University

Evaluation Feedback construct
The evaluation feedback was useful.
The evaluation feedback was timely.
The evaluation feedback was specific.
The evaluation feedback was constructive.
The evaluation feedback helped to improve my instructional effectiveness.
The evaluation feedback represented my instructional ability.
The evaluation feedback informed specific changes in my classroom practice.
The evaluation feedback was aligned with the National Teaching Standards.
The evaluation feedback was aligned with Core Curriculum Content Standards.
The evaluation feedback was aligned with the grade level(s) I teach.
The evaluation feedback was aligned with the subject(s) that I teach.
The evaluation feedback was aligned with the school instructional improvement goals.
The evaluation feedback was aligned with the school district goals.
The evaluation feedback provided information for professional development opportunities.
I was satisfied with the feedback I received from my teacher evaluation.

Teacher Evaluation Experiences Survey – Principal Form (TEES-P) Reddy, Dudek, Kettler, Kurz, & Peters © 2015 Rutgers University

Evaluation Feedback construct
My evaluation feedback was useful.
My evaluation feedback was timely.
My evaluation feedback was specific.
My evaluation feedback was constructive.
My evaluation feedback helped to improve my teachers' instructional effectiveness.
My evaluation feedback represented my teachers' instructional abilities.
My evaluation feedback informed specific changes in my teachers' classroom practices.
My evaluation feedback was aligned with the National Teaching Standards.
My evaluation feedback was aligned with the Core Curriculum Content Standards.
My evaluation feedback was aligned with the grade level(s) taught by my teachers.
My evaluation feedback was aligned with the subject(s) taught by my teachers.
My evaluation feedback was aligned with the school instructional improvement goals.
My evaluation was aligned with the school district goals.
My evaluation feedback provided information for professional development opportunities for my teachers.
I was satisfied with the feedback I provided during my teachers' evaluations.

Appendix D

Rotated Component Matrix for Factor Loadings With Exploratory Factor Analysis With Principal Component Analysis (Principal Responses)

Item	Component 1	Component 2
1. The duration of my classroom observations affected the usefulness of evaluative feedback.		.94
2. The duration of my classroom observations affected the quality of the evaluative feedback.		.90
3. The duration of my classroom observations affected the validity of the observations.		.91
4. My classroom observations were long enough to assess instructional effectiveness.	.72	
5. My classroom observations were long enough to provide useful feedback.	.75	
6. My classroom observations were long enough to provide quality feedback.	.78	
7. My classroom observations were long enough to accurately reflect instructional practices.	.79	
8. Part of an observed lesson represented the quality of the whole lesson.	.60	
9. I would have preferred for my classroom observations to be longer in duration.	-.56	.37
10. I would have preferred shorter observations over longer observations.	.60	-.42

Appendix E

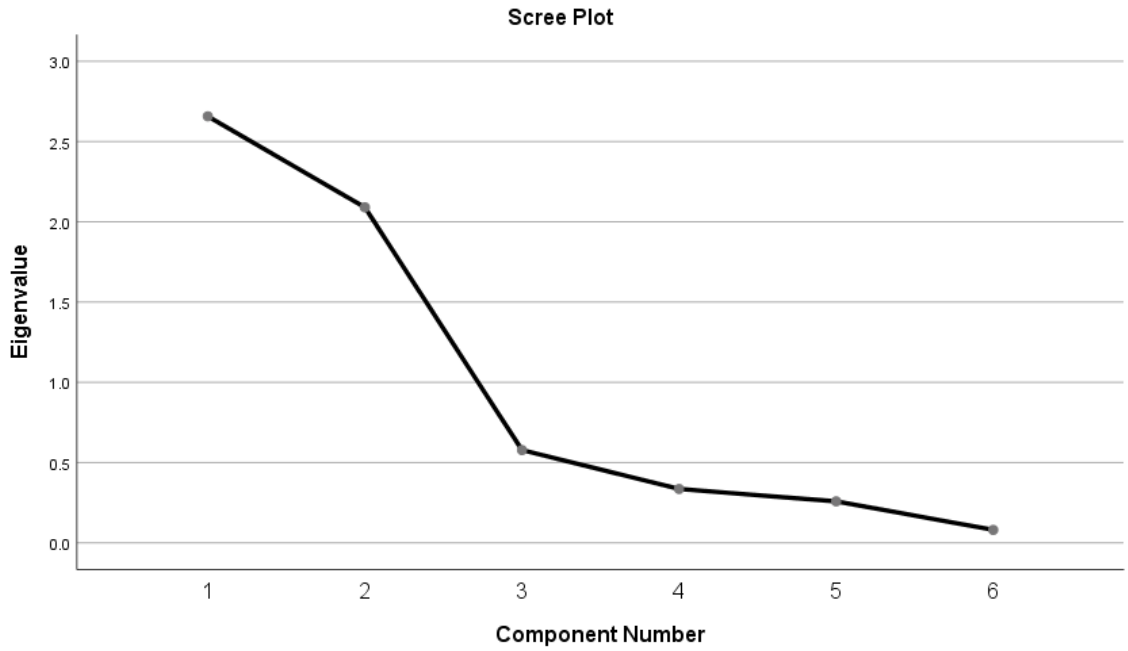
Rotated Component Matrix for Factor Loadings With Exploratory Factor Analysis With Principal Component Analysis (Teacher Responses)

Item	Component 1	Component 2	Component 3
1. The duration of my classroom observations affected the usefulness of evaluative feedback.		.91	
2. The duration of my classroom observations affected the quality of the evaluative feedback.		.88	
3. The duration of my classroom observations affected the validity of the observations.		.77	
4. My classroom observations were long enough to assess instructional effectiveness.	.89		
5. My classroom observations were long enough to provide useful feedback.	.83		
6. My classroom observations were long enough to provide quality feedback.	.87		.34
7. My classroom observations were long enough to accurately reflect instructional practices.	.80		
8. Part of an observed lesson represented the quality of the whole lesson.	.66		
9. I would have preferred for my classroom observations to be longer in duration.	-.44		-.80
10. I would have preferred shorter observations over longer observations.			.93

Appendix F

Rotated Component Matrix for Factor Loadings With Exploratory Factor Analysis With Principal Component Analysis (Principal Responses for Items 1-7)

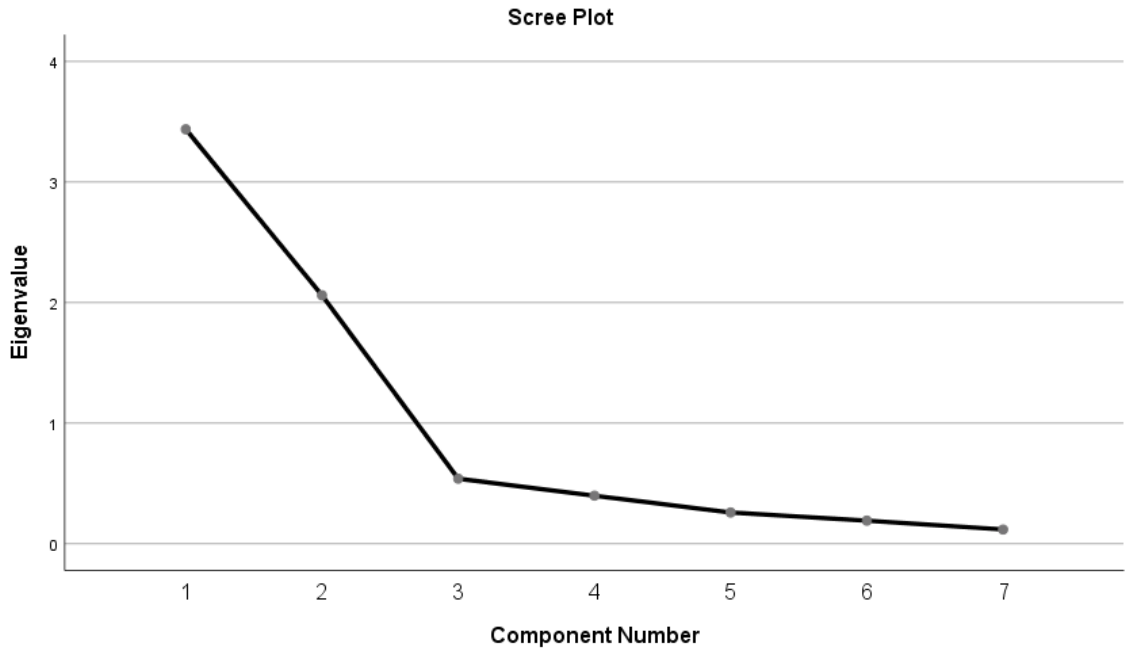
Item	Component 1	Component 2
1. The duration of my classroom observations affected the usefulness of evaluative feedback.	.96	
2. The duration of my classroom observations affected the quality of the evaluative feedback.	.94	
3. The duration of my classroom observations affected the validity of the observations.	.91	
4. My classroom observations were long enough to assess instructional effectiveness.		.79
5. My classroom observations were long enough to provide useful feedback.		.80
6. My classroom observations were long enough to provide quality feedback.		.83
7. My classroom observations were long enough to accurately reflect instructional practices.		.77



Appendix G

Rotated Component Matrix for Factor Loadings With Exploratory Factor Analysis With Principal Component Analysis (Teacher Responses for Items 1-7)

Item	Component 1	Component 2
1. The duration of my classroom observations affected the usefulness of evaluative feedback.		.88
2. The duration of my classroom observations affected the quality of the evaluative feedback.	-.53	.73
3. The duration of my classroom observations affected the validity of the observations.	-.41	.68
4. My classroom observations were long enough to assess instructional effectiveness.	.86	
5. My classroom observations were long enough to provide useful feedback.	.84	
6. My classroom observations were long enough to provide quality feedback.	.91	
7. My classroom observations were long enough to accurately reflect instructional practices.	.80	



Appendix H

Final Survey Instrument

The purpose of this survey is to explore the perceptions of principals and teachers about the duration of classroom observations conducted by principals and evaluative feedback. Participants should give the response that best reflects their own situation and experiences over the past year.

Demographics and Background Information

1. What is your professional role?

    a. Teacher

    b. Principal

2. What is the average duration of each classroom observation you conduct or receive for evaluative purposes?

    a. Less than 15 minutes

    b. 15 – 30 minutes

    c. More than 30 minutes

3. For principals, what is the average number of times each teacher in your school is observed annually for evaluative purposes?

    For teachers, what is the average number of times that you are observed annually for evaluative purposes?

    a. 1 – 3

    b. 4 – 6

    c. 7 – 9

d. 10 or more

Participants will respond to Items 4 – 19 on a Likert scale:

1 – strongly disagree

2 – disagree

3 – neutral

4 – agree

5 – strongly agree

Duration of Classroom Observations

4. The duration of my classroom observations affected the usefulness of evaluative feedback.
5. The duration of my classroom observations affected the quality of the evaluative feedback.
6. The duration of my classroom observations affected the validity of the observations.
7. My classroom observations were long enough to assess instructional effectiveness.
8. My classroom observations were long enough to provide useful feedback.
9. My classroom observations were long enough to provide quality feedback.
10. My classroom observations were long enough to accurately reflect instructional practices.

Evaluation Feedback (TEES-T)

11. The evaluation feedback was useful.
12. The evaluation feedback was timely.

13. The evaluation feedback was specific.
14. The evaluation feedback was constructive.
15. The evaluation feedback helped to improve my instructional effectiveness.
16. The evaluation feedback represented my instructional ability.
17. The evaluation feedback informed specific changes in my classroom practices.
18. The evaluation feedback was aligned with grade level(s) I teach.
19. The evaluation feedback was aligned with the subject(s) that I teach.
20. The evaluation feedback was aligned with the school instructional improvement goals.
21. The evaluation feedback was aligned with the school district goals.
22. The evaluation feedback provided information for professional development opportunities.
23. I was satisfied with the feedback I received from my teacher evaluations.

Evaluation Feedback (TEES-P)

11. My evaluation feedback was useful.
12. My evaluation feedback was timely.
13. My evaluation feedback was specific.
14. My evaluation feedback was constructive.
15. My evaluation feedback helped to improve my teachers' instructional effectiveness.
16. My evaluation feedback represented my teachers' instructional ability.

17. My evaluation feedback informed specific changes in my teachers' classroom practices.
18. My evaluation feedback was aligned with the grade level(s) taught by my teachers.
19. My evaluation feedback was aligned with the subject(s) taught by my teachers.
20. My evaluation feedback was aligned with the school instructional improvement goals.
21. My evaluation feedback was aligned with the school district goals.
22. My evaluation feedback provided information for professional development opportunities for my teachers.
23. I was satisfied with the feedback I provided during my teachers' evaluations.

Appendix I

Recruitment E-mail

Dear Secondary Principals and Teachers,

I am writing to request your participation in a survey to measure principal and teacher perceptions about the duration of classroom observations and evaluative feedback. I am a Doctoral student at Southwest Baptist University in Bolivar, Missouri and a practicing educator. Completing the survey should take less than 5 minutes.

This project has been reviewed by the Southwest Baptist University Research Review Board for research and research-related activities involving human subjects. I would be glad to share the completed study with you upon request.

The survey is anonymous and no identifying information is collected. The completed survey serves as your implied consent to be surveyed. Your participation is voluntary. Should you choose to participate, you are not required to respond to all survey items. You may discontinue completion of the survey at any time. Your survey responses will be strictly confidential and data from this research will be reported only in the aggregate. There are no foreseeable risks associated with participation in this study.

This message is being sent to principals. If you consent to participation, I request that the message be forwarded to the teachers in your school with one year of experience or more for their consideration. To begin the brief survey, click on the following link: https://www.questionpro.com/t/AMyjHZcLCA. I thank you, in advance, for your time and contribution to this research project.

Sincerely,

David Pyle

Appendix J

Final Survey Scree Plots