Effectiveness of Principals as Evaluators of Teachers

A PROJECT
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Shirley Ann Gregoire

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF EDUCATION

Neal C. Nickerson, Jr., Ed.D.

November 2009

Acknowledgements and Dedications

This project would never have been completed without the support, inspiration, and encouragement given to me by friends and family. Throughout my work on this project I always felt the echoes of encouragement from my mother, Lenora Horinek Gregoire, who passed away in 1983. She was a woman before her time who applied to be an Air Force pilot in the 1940's. She always stood up for her daughters telling us we were just as important and significant as "the boys." I also know it was the drive, tenacity, and value of hard work I acquired from my dad, Giles Gregoire, that carried me forward to completion.

I dedicate this book to my sons, Scott Kratochvil and Ryan Kratochvil. Their love and encouragement fueled my persistence. I am so blessed to have such wonderful sons who have always been there for me.

I thank Patricia Gregoire, Vernon Gregoire, Karen Myers, Mary Jean Gregoire, and Wayne Gregoire - my sisters and brothers - for their support from afar.

Finally, I thank Neal Nickerson, my esteemed advisor, for all his patience and guidance. Memories of our lunches at White Castle will always be with me.

Abstract

The purpose of this study was to gather perceptions of principals and teachers with regard to the effectiveness of principals as evaluators of teachers. Perceptions were reviewed within the context of seven standards across the four attributes of the personnel standards developed by the Joint Committee on Standards for Educational Evaluation (Joint Committee, 2009). These standards, organized by attributes, define quality personnel evaluation in education.

The research was exploratory and utilized quantitative research methods. Principal and teacher participants were asked to identify their perceptions of the frequency with which principals followed effective evaluation practices. Teacher participants were also asked to identify the relative importance of the practices as a factor of effective evaluation. Because validity and reliability are of heightened importance when evaluation results are applied to performance pay decisions (Loup & Ellett, 1997), the study used perceptions of middle school principals and tenured teachers who participated in the Minnesota Q-Comp program (Q-Comp), a performance based merit pay system.

An analysis of the findings indicated that principals are effective evaluators of teachers. Teachers rated principals as “Often” following the effective practices on 16 of the 24 practices. Principals’ self ratings generally mirrored those of the teachers with principals frequently rating themselves higher. A further analysis based on the relative importance of each practice as ranked by teachers further supports principal effectiveness as evaluators. Principals’ practices which are ranked higher in frequency are the practices

that teachers identified as most important. Principals' practices that are ranked lower in frequency are of lower importance to the teachers.

This study found that principals are effective evaluators of tenured teachers as determined by the frequency with which they follow national personnel evaluation standards. The study suggests implications for principal preparation programs and district in-service training based on the effective practice skills analysis. The study further suggests that principals are capable of assuming a significant role in tenured teacher evaluation in a performance pay system.

TABLE OF CONTENTS

LIST OF TABLES ..... **vii**

LIST OF FIGURES ..... **vii**

CHAPTER I: INTRODUCTION **1**

Research Statement 2
Background of the Problem 4
Significance and Rationale of the Study 7
Research Questions 9
Limitations of the Study 10
Definition of Terms 10

CHAPTER II: REVIEW OF THE LITERATURE **12**

Introduction 12
Personnel Evaluation Standards 25
Approaches to Evaluation - Models 31
Principals and Teacher Evaluation 46
Teacher Performance Pay and Teacher Evaluation 56
Summary 60

CHAPTER III: METHODOLOGY **62**

Introduction 62
Research Design 63
Research Population 63
Sample Selection 64
Research Instruments 67
Development of the Research Instruments 68
Method of Data Collection 70

CHAPTER IV: RESULTS **72**

Introduction 72
Demographic Data Analysis 72
Statistical Analysis by Effective Practices 74
Statistical Analysis by Attribute 80

CHAPTER V: DISCUSSION **85**

Introduction 85
Summary of Purpose 85
Significance and Rationale of the Study 86
Review of Procedures 87
Review of Main Findings 87
Conclusion 93
Recommendations 94
Suggestions for Further Research 95

REFERENCES: 96
APPENDIX A: LIST OF QUALIFYING SCHOOLS 113
APPENDIX B: SURVEYS 116
APPENDIX C: CONSENT FORMS 121
APPENDIX D: **STATISTICAL TABLE 125**

LIST OF TABLES

Table 1	Population Sample of Minnesota Q-Comp Middle Schools	64
Table 2	Population Sample of Teachers and Principals	65
Table 3	Summary of the Response Rate of Study Participants	66
Table 4	Relationship of Survey Scale Statements to Personnel Evaluation Standards	67
Table 5	Characteristics of Teacher Participants	73
Table 6	Characteristics of Principal Participants	74
Table 7	Average Score and Ranking of Each Statement	76
Table 8	Frequency of Effective Practices between Teachers and Principals	77
Table 9	Average Score and Ranking of Statements	78

LIST OF FIGURES

Figure 1 Average Frequency and Importance Scores 80

CHAPTER I

INTRODUCTION

Historically, teacher evaluations have often been regarded as meaningless, bureaucratic rituals by both teachers and administrators (Wise, Darling-Hammond, McLaughlin, & Bernstein, 1984b). While evaluations of non-tenured teachers give direction to contract renewal decisions, evaluations of tenured teachers seem purposeless given continuing contract statutes. Thus, principals tend to complete evaluations as directed by district policy in a perfunctory manner. As the accountability era of the 1990's and 2000's passed through education systems, the focus of teacher evaluation shifted from being a tool for accountability to being a tool for professional development. Principals were called upon to be instructional leaders guiding teachers in curriculum and instruction. Skeptics questioned the appropriateness of principals as evaluators in this new era by citing lack of subject content knowledge and distance from contemporary classroom operations. Peer mentors and coaches from the teaching ranks were brought into the role of evaluators in response to the new professional development focus of evaluation. This trend brought to the forefront a questioning of the effectiveness of principals as evaluators of teachers.

Determining the effectiveness of the principal as an evaluator of teachers is not merely a matter of checking to insure evaluation forms are filed correctly according to policy. The effectiveness answer lies within the ebb and flow of teacher evaluation purposes and processes. Principals' effectiveness in evaluation differs depending upon the purpose of the evaluation with some purposes better met than others (Bridges, 1990). However, evaluation systems generally serve not one but several purposes (Wise et al.,

1984b) thus adding complexity to the issue. Controversy also surrounds the data used for evaluation. Evaluation data sources such as achievement data, classroom observations, and student or parent surveys vary in validity and worth. While principals typically do not have authority to choose which data are used, their ability to gather, analyze, and interpret data may be different dependent on the data source. Finally, teacher evaluation is politically challenging to principals. While principals must have integrity in performing teacher evaluations, they also need to maintain positive working relationships with their staffs in order to perform their leadership roles. Thus, the general effectiveness of principals as evaluators of teachers is open to professional scrutiny and difficult to determine.

In spite of the challenges of teacher evaluation, principals are typically responsible for both tenured and non-tenured teacher evaluation. This study seeks to inform principals, districts, and principal preparation programs of the strengths and deficiencies of principals as evaluators. The study uses nationally accepted standards for evaluation as the lens to make determinations. Chapter 1 introduces the research statement, background information, research questions, significance and rationale, limitations of the study, and definition of terms.

Research Statement

The primary purpose of this study is to gather perceptions of principals and teachers with regard to the effectiveness of principals as evaluators of teachers. Perceptions are viewed within the context of personnel evaluation standards developed by the Joint Committee on Standards for Educational Evaluation (Joint Committee, 2009). These standards, organized by attributes, define quality education personnel

evaluation. Furthermore, the study narrowed the research population to include only teachers and principals whose schools participated in the Minnesota Quality Compensation for Teacher Program (Q-Comp), a performance based merit pay program. This target population was selected based on research which suggests that teachers involved in performance pay compensation systems place more importance on evaluation decisions (Jacob & Lefgren, 2008). The target population was further narrowed to include only tenured teachers due to their more extensive experiences with principals as evaluators. Thus, the study uses perceptions of middle school principals and tenured teachers who participated in the Minnesota Q-Comp program in 2007-2008 to determine the effectiveness of principals as evaluators of teachers through determining the perceived strengths and deficiencies of their evaluation practices as defined by nationally accepted evaluation standards.

For purposes of this study, performance pay is defined as awarding compensation based upon a teacher's instructional skills and contributions to the professional community while merit pay bases compensation on a teacher's instructional skills and contributions and, additionally, on the achievement of identified outcomes. The Minnesota Q-Comp program allows districts adopting the program locally to define process but dictates that it must include some combination of classroom performance, outcomes, and student achievement in determining the awarding of pay. Currently, principals in some Q-Comp districts influence the awarding of performance pay while others do not thus underscoring the inconsistent perspective on the role of the principal.

In this study, effective evaluators are defined as those evaluators that most closely follow practices which meet the Personnel Evaluation Standards developed by the Joint

Committee on Standards for Educational Evaluation (Joint Committee, 2009). Recognizing that not all traits are equally valued, the study also identifies those traits which most significantly contribute to an evaluator's effectiveness as determined by teachers.

Background of the Problem

Teacher evaluation has evolved over the past 80 years as political and societal pressures recast American education. Educational trends have shaped and reshaped the purpose of evaluation resulting in subsequent changes in models, processes, and data. Within the field of evaluation, a model is defined as the overarching framework which supports the stated purpose and philosophy. A process addresses the "who," "what," and "when." Data are the evidence which is gathered, analyzed, and drives decisions within the system. The evaluator is one component of the process. While principals have historically served as evaluators, recent redefinitions of the purposes and consequences of teacher evaluation have ignited controversy surrounding this role (Peterson, 2000).

Evaluators have been viewed differently over the years. In the "Inspector Model" of the early 1900's (Scriven, 1995b), the evaluator's primary responsibility was to control the behaviors of the teachers, insure the classes were well managed, and make sure the school was cared for properly. Evaluators were considered experts who commended excellence. This bureaucratic approach continued into the 1960's as evaluators required teachers to comply with the use of the latest educational techniques and methodologies. A new wave of teacher evaluation approaches began in the 1980's with the advent of the professional movement. Schon (1983) expanded on the use of reflective practice as a means of solving problems and improving instruction. From Schon's work, Goldhammer

and Cogan developed the clinical supervision model which focused on data collection, instructional activities, and frequent feedback through multiple evaluations. Glickman further developed this model through differentiating evaluation based upon a teacher's developmental level and stage of development. Scriven classified these models as method models with their focus on methodology and reflection rather than firm criterion or outcomes. As the evaluator's role moved from that of an expert judge towards professional coach, staff other than principals became involved in teacher evaluation (Danielson & McGreal, 2000).

At the same time that teacher evaluation shifted to a reflective practice model, corporate America became more accountability and results based. Earlier studies highlighting the deficiencies in American education (*A Nation at Risk,* National Commission on Excellence in Education, 1983; *What Matters Most: Teaching for America's Future*, National Commission on Teaching and America's Future, 1996) generated scholarly work which sought to identify the traits of effective teachers. These studies were followed by the No Child Left Behind (NCLB) legislation enacted in 2001 which set into place an accountability system for student achievement. Part of the legislation recommended performance pay for teachers based on merit. The Minnesota legislature responded by developing the Q-Comp program in 2005 which awarded performance pay based on student achievement results, outcomes, and teacher performance as measured by evaluations. While the legislation mandated the use of peer coaches as evaluators, the role of the principal was left unaddressed with principal evaluations incorporated into some, but not all, districts' performance pay equations. For those participating Q-Comp districts, evaluation of tenured teachers now had more wide

reaching consequences. Accordingly, the effectiveness of principals' evaluations took on a new relevance in the face of high stakes evaluation.

Research strongly supports the importance of the principal in teacher evaluation (Colby, Bradshaw, & Joyner, 2002). Indeed, principals play an essential role in determining the attitude, value, and effectiveness of teacher evaluation within a building (Davis, Ellett, & Annunziata, 2002). Despite the key role they play, principals face many challenges in their roles as evaluators. Prescribed checklists and other district mandated forms can limit reports. Studies have indicated biased reporting, unrepresentative sampling, and limited number of samplings (Peterson, 2004; Scriven, 1981). Furthermore, principals by the nature of their leadership role must support collaboration within their schools. The passing of judgment on the teaching quality can cause sociological conflicts (Wise et al., 1984b). Technically accurate tools for data gathering and analysis are often unavailable (Peterson, 2000; Scriven, 1981). Principals do not have the time to do teacher evaluation with the frequency and depth that is required (K. Peterson & C. Peterson, 2006). Faced with these challenges, alternative personnel such as mentor teachers or coaches have taken on summative and formative evaluator roles under the assumption that they can better help teachers improve their craft.

Controversy regarding the effectiveness of evaluation was the impetus that led to the formation of the Joint Committee on Standards for Educational Evaluation. The committee was organized in 1975 with the purpose of providing a credible, systematic mechanism for examining evaluation practice (Joint Committee, 1988). Driven by documented deficiencies in personnel evaluation in education, the Joint Committee undertook the development of personnel evaluation standards for education (Joint

Committee, 1988). The process of developing the standards for the first edition involved extensive input and review at a national level. The second revision of the standards was published in 2009 following a review of the standards by the Joint Committee whose members then represented 15 major education organizations.

The Personnel Evaluation Standards provide a logical framework for research in the area of effective evaluation. By design, the standards can be applied to a wide range of evaluation techniques including observation, interview, development of a portfolio, student assessment, and goal development. They do not attempt to define good teaching but rather good evaluation. The standards' assumptions recognize that the users of the standards must define their own educational goals, unique situations, and apply the standards accordingly. Finally, the Joint Committee explicitly enumerated the uses of the standards including "as a logical structure for deriving and investigating questions and hypotheses about personnel evaluation" (Joint Committee, 1988, p. 15).

The principal's role as an evaluator now stands at a crossroad. The principal's effectiveness as an evaluator is challenged on the basis of depth of specific knowledge and motivation. Yet research studies have shown them to be effective judges of high and low performing teachers (Jacob & Lefgren, 2005). The standards provide a means of giving direction to the question of the effectiveness of principals as evaluators of teachers and are the basis of this study.

Significance and Rationale of the Study

Several aspects of the study are unique. First, the data used to determine the perceived effectiveness of principals as evaluators is based on nationally established standards for personnel evaluation. The study attempts to determine effectiveness based

on the degree to which principals follow the standards' evaluation practices when evaluating teachers. The literature contains numerous examples of studies involving teacher evaluation by principals based on other types of factors. Comparative studies have examined correlations between teacher rating by principals and student, parent, peer, or self ratings (Peterson, Wahlquist, & Bone, 2000; Peterson, Wahlquist, Brown, & Mukhopadhyay, 2003; Wilkerson, Manatt, Rogers, & Maughan, 2000). Other studies have compared teacher ratings by principals to student achievement (Jacob & Lefgren, 2008; Medley & Coker, 1987). However, no study could be located that based evaluation effectiveness on the frequency by which standards of evaluative practice are followed. Furthermore, the study also provides data on the relative importance of each practice as perceived by the teachers, thus allowing for an effectiveness analysis based on not only the frequency by which a practice is followed but also on the relative importance of that practice as an element of effective evaluation.

Secondly, the study focuses on tenured teacher evaluation. Tenured teachers are much more skeptical of personnel evaluations than are non-tenured teachers (Kauchak, Peterson, & Driscoll, 1985). Conversely, the need for effective tenured teacher evaluation is evident in empirical studies which indicate that while 5-15% of the teachers in public school classrooms perform at incompetent levels most tenured teachers receive outstanding scores (Tucker, 1997). Thus, while the need for effective evaluation of tenured teachers is clear in the literature, few studies address only tenured teacher evaluation.

Finally, there is little research on the effectiveness of principals' evaluation practices using data gathered within the context of a performance pay system such as the

Minnesota Q-Comp program. In these programs, literature suggests the competency of the evaluator and the resulting validity of the evaluation become more critical as the evaluation becomes high stakes (Loup & Ellett, 1997; Loup, Garland, Ellett, & Rugutt 1996).

Results of the study will inform principal practice. The findings will not only inform principal professional organizations, districts, and principal preparation programs of the strengths and weaknesses of principals' teacher evaluation practices, but will also shed light on the viability of utilizing principals as primary decision makers in awarding performance pay.

Research Questions

The primary purpose of this study is to gather perceptions of principals and teachers with regard to the effectiveness of principals as evaluators of teachers. The effectiveness of the evaluations is based on the frequency by which the principals' evaluative practices follow practices outlined in seven of the personnel standards developed by the Joint Committee on Standards for Educational Evaluation (Joint Committee, 2009). These standards, organized by attributes, define quality education personnel evaluation. In addition, teacher participants are also asked to identify the relative importance of the practices as a factor of effective evaluation. The study uses the perceptions of middle school principals and tenured teachers who participated in the Minnesota Q-Comp program in 2007-2008 to answer the following research questions:

1. Which practices of effective evaluators do principals follow as perceived by teachers and principals?
2. Which practices of effective evaluation are most important to teachers?
3. What is the relationship between the effective practices that principals follow and those that are most important to the teachers?

Related goals of the study include implications of the findings for district, organizational, and principal preparation programs; and the viability of using principals as primary decision makers in awarding performance pay.

Limitations of the Study

The following limitations are noted:

1. The study is limited to middle school teachers and principals and is not applicable to elementary or high school teachers and principals.
2. The study is limited to tenured teachers and is not applicable to non-tenured teachers.
3. The study is limited to examining evaluation as a means of determining competency and is not applicable to evaluation used to improve performance.

Definition of Terms

Certain terms used in this study are defined to clarify their meanings in the study:

1. Effective teacher evaluation: Teacher evaluation conducted in a manner that follows the practices set forth in the Personnel Evaluation Standards (Joint Committee, 2009).
2. Evaluation: The systematic assessment of a person's performance and/or qualifications in relation to a professional role and some specific and defensible institutional purpose (Joint Committee, 1988, p.7).
3. Formative evaluation: An evaluation that is the basis for professional development.
4. Merit pay: A system for awarding compensation based on a teacher's instructional skills, contributions, and on the achievement of identified outcomes.
5. Models of teacher evaluation: Schematic description of a personnel evaluation system that accounts for its known or inferred properties including methodology, evaluative criteria, and procedures.
6. Performance pay: A system for awarding compensation based upon a teacher's instructional skills and contributions to the professional community.
7. Principal: The chief administrator on a school campus who has responsibility for the total school program.

8. Standards: A principle commonly agreed to by people engaged in the professional practice of evaluation for the measurement of the value or the quality of an evaluation (Joint Committee, 1981, p.12).

9. Summative evaluation: An evaluation that is a judgment of quality of work.

10. Supervision: The process of observing and dialoguing with a teacher for the purpose of assisting the teacher in improving practice.

CHAPTER II

REVIEW OF THE LITERATURE

Introduction

While other industries have clarified and supported the role of the supervisor in determining rewards and sanctions for employees, this has not been true for principals within the field of education. In 2001 the No Child Left Behind Act (NCLB) - supported by *Actions for Excellence* (1983) and *A Nation At Risk* (1983) - highlighted teacher quality as a critical element in the success of students. These reports brought teacher quality and accountability to the forefront of educational issues. As a result, several states, including Minnesota, initiated legislation to support alternative pay programs for compensation based upon merit. In these programs, the competency of the evaluator, the clarity of the evaluative criteria, and the resulting validity of the evaluation becomes more critical. The obligations of the principal have generally been defined in the area of evaluation and termination through state statute (Holland, 2001). The principal's role is undefined in state statute referencing Minnesota teacher alternative pay programs (Minn. Stat. 122A.413-415). Research findings are also inconclusive as to the effectiveness of principals as evaluators of teachers. Chapter 2 investigates the various facets of teacher evaluation and the principal's role within that context. Chapter 2 is divided into the following five sections:

Section 1: Background of the Problem

Section 2: Personnel Evaluation Standards

Section 3: Approaches to Evaluation - Models

Section 4: Principals and Teacher Evaluation

Section 5: Teacher Performance Pay and Teacher Evaluation

The first section is a review of literature on the various theories and educational initiatives that have shaped teacher evaluation and the principal's role within it. The discussion is presented through four different topics: (a) historical context, (b) significance of evaluation, (c) purpose of evaluation, and (d) issues in evaluation.

The second section reviews the development of the standards for evaluation of education personnel. These nationally recognized standards describe quality evaluation systems and practices and are the basis of this study. This section examines the background of the project, general attributes, and application of the standards.

The third section reviews current literature on the various approaches to teacher evaluation. These approaches are framed by various theoretical models of teacher evaluation. The models are divided into two categories: criteria based and process based. Literature indicates that the conceptual framework of various teacher evaluation models impacts the role and effectiveness of the principal as an evaluator of teachers.

The fourth section is a review of the literature of principals as evaluators of teachers. Both the technical aspects and the sociological aspects of conducting valid evaluations are examined. The section concludes with an analysis of themes of agreement and disagreement within the literature. The review is limited to effective evaluations as a means of quantifying the performance of a teacher not as a means of effecting a change in practice.

The last section examines performance pay systems for teachers. Evaluations performed within performance pay systems are higher stakes thus driving a need for

greater accuracy on the part of the evaluator (Loup & Ellett, 1997). The rationale behind the development of the systems and related implications for principals are scrutinized.

Background of the Problem

The nature of teacher evaluation has changed over the past century in response to accountability measures within K-12 education, research on the relationship between teacher quality and student achievement, change theories of adult behavior, and data based decision making. The *Nation at Risk* Report (1983) signaled a fundamental shift in K-12 American education. The recommendations from this report generated a string of effective teaching and accountability studies. The studies affirmed the importance of having quality teachers in the nation's classrooms as a means of increasing student achievement. Educational change literature examined the effectiveness of professional development based on reflection and dialogue. A shift in the purpose of personnel evaluation from judgment to improvement caused the traditional principal's evaluation to be re-examined. As the evaluation field became embroiled in controversy, the principal's role became less clear.

The ambiguity of the role of the principal is the result of multiple dynamics. The literature provides a spectrum of positions on the role of the principal varying from insignificant to critical dependent upon the context, significance, and purpose of personnel evaluation. In this section these variables will be examined historically followed by a review of the literature on current issues in evaluation.

Historical Context

The role of the principal in evaluation has been influenced by social, political, and technical shifts in education for over a hundred years. In the early 1900's education was viewed primarily as a means of social control preparing children to blend into American

society. Most recently, education has been influenced by business models which use statistical results to determine success or failure. Seemingly in conflict with a hard data approach, the professional approach to evaluation suggests the use of self reflection or peer dialogue to meet an evaluative goal of improving practice. These two varying perspectives place principals in very different roles. Are principals guides on the side or sages on the stage? Can principals effectively fulfill either of those roles? The development of the opposing roles will be examined in light of three movements in evaluation which have emerged over the last century: the professional, effective teaching, and results-based movements.

The professionalism movement is characterized by a non-directive approach to supervision. Supervision from this perspective is generally defined as the process of engaging teachers in professional dialogue and reflection for the purposes of improving practice and increasing student achievement (Zepeda, 2003). The main focus of supervision is not to change behavior but rather to engage the teacher successfully in reflective practice (Siens, 1996). In contrast to supervision, evaluation is associated with determining the worth of a teacher through the gathering of various types of supportive data (Darling-Hammond, 1989). Typically, evaluation is used for accountability purposes and personnel decisions. Elements of both supervision and evaluation become blended in various evaluation models lending to the confusion regarding the principal's role in either area. Specifically, the literature does not agree on the role of the principal within the professional movement. Some evidence indicates principals to be ineffective and other evidence indicates the contrary (Wise et al., 1984b).

The professional movement is rooted in the work of Schon (1983) and Little's (1982) collegiality studies. Evidence from these studies suggested that more change in instructional practices occurs when teachers have the opportunity to reflect upon their practice with a colleague or peer. In a hallmark study funded by the Rand Corporation, Wise et al. (1984b) recommended that teachers be regarded as professionals and encouraged professional dialogue. Costa (1984), Wise et al., and Danielson and McGreal's (2000) work strengthened support for developing teacher teams to concentrate on continuous improvement. The professional movement focused on involving teachers in critiquing their work. The trend towards professionalism continued into the 90's with the advent of cognitive coaching (Costa & Garmston, 1994) and peer coaching (Showers, 1984). Research indicated that teachers perceived that the traditional evaluation methodologies based upon the clinical supervision model (Acheson & Gall, 1980) were ineffective in assisting them in improving practice (Wise et al., 1984b). Thus, two forms of evaluation emerged with two differing purposes: (a) formative evaluation in which the goal is to enable a teacher to identify their own strengths and weakness and plan appropriate professional development activities and (b) summative evaluation which provides a basis for administration decision making with regards to personnel decisions including hiring, firing, and tenure. Formative evaluation became more closely associated with the role of a supervisor; summative evaluation became more closely associated with the role of an evaluator. Thus, the professional movement divided the field of personnel evaluation leaving the role of the principal within each field undefined.

The effective teaching approach to evaluation compares a teacher's practices to those which, according to research, result in improved student achievement. Good

teaching practices were identified through the work of the School Improvement Model Project (Manatt, 1994). Later, these practices were used to develop extensive job descriptions for teachers (see Manatt). Further definition was given to quality teaching through the National Professional Standards Board which was established under the auspices of the Carnegie Foundation in 1987 with the charge of defining a highly professional teacher. The Interstate New Teacher Assessment and Support Consortium (INTASC) also developed a parallel set of standards for new teachers. In 1998 North Carolina adopted the new standards for teaching as the basis for teacher evaluations. Principals have varying roles within the effective teaching approach. Both formative and summative evaluations using this approach can include principal, peer, and coach involvement.

The results-based approach relies on the gathering of objective data to provide evidence of student achievement. The focus is not on reflection or improvement in practice. Information about student achievement based on standardized test results was used to determine teacher proficiency in the early 80's (Millman, 1981); however, researchers challenged this use based on validity and reliability (Medley & Coker, 1987). The concept of marrying descriptors of the duties of a teacher with multiple sources of evidence was advanced by two significant theoretical models: Judgment Based Teacher Evaluation (Popham, 1988b) and Duty Based Teacher Evaluation (Scriven, 1988). Both of these models were summative with an administrator or team of reviewers examining the evidence to determine the worth of the teacher. This approach was overshadowed by the professional movement of the 90's.

Currently, teacher evaluation is in a state of flux. The influences of all three movements result in very disparate philosophies and models. The role of the principal within this confluence is also in flux. Legally, principals are authorized by the state of Minnesota to evaluate teachers (MN Administrative Rules 2007 Part 3512.0300). Furthermore, school districts are required by statute to provide written evaluations three times per year for all non-tenured staff who have performed services for 120 days or more, two times per year for non-tenured teachers performing services from 60-119 days and one time per year for non-tenured teachers performing service for less than 60 days per year (MN Statute Chapter 122.A40, 2007). Minnesota statute, however, only designates the evaluations be completed. Minnesota statute directs the school board (MN Statute Chapter 122.A40, 2007) or the school site management team (MN Statute Chapter 122.A41, 2007) to determine who performs the evaluations. For continuing contract teachers, Minnesota statute directs the school board and an exclusive representative of the teachers to develop a peer review process. Thus, for all Minnesota school districts, the legitimate authority for principals to make decisions regarding compensation does not exist.

The Significance of Evaluation

Research studies are inconclusive regarding teachers' perceptions of the importance of evaluation by principals. While teachers noted that principal evaluations are valuable as a means of obtaining appreciation for their work and improved communication, they also noted personnel evaluations by principals are not valuable as a means to inform practice (Wise et al., 1984b). These views drive the need for an

explanation of both the rationale and effectiveness of evaluating teachers from a summative and formative perspective.

The rationale for evaluating teachers is influenced by the current political drive for accountability and by research findings which substantiate the connection between teacher quality and student achievement. Politically, the nation is living in an era of accountability. Society has become consumer and results orientated. The public no longer takes institutions at their word but rather requires data and credible evidence for results. In education teacher quality data is most commonly gathered through the teacher evaluation process (Bridges, 1992). Student achievement data is most commonly gathered through student achievement tests.

Numerous studies have attempted to link teacher quality to student achievement test scores. In 1984, a significant study of teacher evaluation (Wise et al., 1984b) suggested that student achievement cannot be raised without improving the quality of teachers that are teaching the students. While Medley and Coker's 1987 study disputed the correlation between student achievement and teacher quality, advances in statistical testing models promoted the theory that teachers add value to the achievement level of students in a manner that is long term and cumulative (Sanders & Horn 1998; Sanders & Rivers 1996). Indeed, a 2003 study conducted by RAND Education (McCaffrey, Lockwood, Koretz, Louis, & Hamilton) reviewed the major studies in the area of teacher effect and concluded that teachers differentially affect student achievement. Thus, the political climate coupled with mounting evidence of the significance of quality teachers has propelled teacher evaluation back into the limelight.

While the rationale for evaluation is apparent in the literature, the evidence regarding the effectiveness of teacher evaluation is mixed. Validation of effectiveness varies based upon whether the district's stated purpose for teacher evaluation is accountability for quality or improvement in practice. Due to various factors, including the constraints of teacher tenure laws, limited administrative expertise in content areas, and unclear or inappropriate criteria, administrators and teachers have challenged the effectiveness of teacher evaluation concluding the process is meaningless (see Danielson & McGreal, 2000) or a waste of time (Holland, 2005).

Effectiveness of evaluation varies depending upon the purpose of the evaluation. When teacher evaluation focuses on improvement in practice, evidence indicates that teacher-specific classroom behavior or organization is influenced, but not content-related instructional practices (Kimball, 2002). Thus, theorists speculate that evaluations by coaches, peer mentors, or subject area specialists may be more effective (Wise et al., 1984b). When teacher evaluation focuses on accountability measures, research indicates that evaluations are accurate for low performing and high performing teachers while less accurate for teachers in the middle of the spectrum (Jacob & Lefgren, 2008). Conversely, earlier studies identified a low level of correlation between principals' judgments and teacher effectiveness (Medley & Coker, 1987). In spite of frustration with evaluation systems, administrators do place high value on identifying and removing incompetent teachers (Bridges, 1992). The greater importance a district places on the use of evaluation, the greater the willingness of a principal to confront incompetent teachers (Bridges 1992). Principals do feel they have the skills and ability to accomplish this goal (Painter, 2000). Indeed, teacher evaluation by administrators, while not highly accurate,

may be just as effective as other means of providing accountability for such actions as performance pay and personnel actions (Jacob & Lefgren, 2005).

The Purpose of Evaluation

The importance of identifying the purpose of evaluation cannot be overstated. The purpose is the reason why an evaluation system is put into place. In order for the evaluation to be valid, the purpose must define the process while the process will give validity to the results. The hallmark study conducted by Rand Education for the Carnegie Corporation (Wise et al., 1984b) recommended that a district should change the process if the purpose changes. Initially, the primary purpose of teacher evaluation was quality assurance (e.g., see Danielson, 2001). Over the past 25 years, theorists have expanded upon the purposes of evaluation. A 10 year follow-up study to a 1985 study of the 100 largest school districts in the United States (Ellett & Garland, 1987) confirmed that the main purpose of evaluation was still personnel decisions (Loup, Garland, Ellett, & Rugutt, 1996). However, the study also noted a larger portion of survey respondents reported improvement in practice as a primary focus. The trend towards a purpose of improvement, either individual or school improvement, has continued for the subsequent ten years (Gordon, 2006; Mathers, Oliva, & Laine, 2008).

Historically, educational researchers have categorized the purposes of evaluation in numerous ways. Wise et al., (1984b) identified four basic purposes: (a) individual improvement, (b) individual accountability, (c) organizational improvement, and (d) organizational accountability. Danielson and McGreal (2000) noted two primary purposes for teacher evaluation: quality assurance and professional development. Stronge (in Stronge, 2006) defined two types of evaluation based on Danielson and McGreal's

purposes: accountability oriented and improvement oriented evaluation. Accountability oriented evaluation is generally associated with summative evaluation while improvement orientated evaluation is associated with formative evaluation.

Summative evaluation calls for a judgment which results in decisions regarding promotion, retention, tenure, merit pay, and other personnel decisions. Because of the high stakes nature of summative evaluation, the process must yield legally defensible evidence. Formative assessment encourages teacher development and the improvement of instructional quality. A formative evaluation process provides direction for staff development, recognizes and reinforces outstanding practice, and provides constructive feedback for improved instruction. Because of the focus of formative evaluation, the evaluator must be knowledgeable and trusted in a specific instructional area. Which purpose is most important? Can both purposes be blended together in one evaluation system? This topic has been at the forefront of much educational debate.

Issues in Evaluation

The pivotal study, *Teacher Evaluation: A Study of Effective Practices* (Wise et al., 1984b) which was conducted by RAND Corporation for the National Institute of Education, set the stage for a series of studies regarding teacher evaluation practices. The study cited five conclusions that were necessary conditions for successful teacher evaluation:

1. To succeed, a teacher evaluation system must suit the educational goals, management style, and conception of teaching and community values of the school district.
2. Top-level commitment to and resources for evaluation outweigh checklists and procedures.
3. The school district should decide the main purpose of its teacher evaluation system and then match the process to the purpose.

4. To sustain resource commitments and political support, teacher evaluation must be seen to have utility. Utility depends on the efficient use of resources to achieve reliability, validity, and cost-effectiveness.

5. Teacher involvement and responsibility improve the quality of teacher evaluation.

Based upon data collected in 1985, Ellett and Garland (1987) reported the results of a national survey of teacher evaluation practices from the 100 largest school districts in the United States. Key findings of the study pointed to concerns about instruments, practices, and policies. In a replication of the study 10 years later, Loup et al. (1996) concluded that practices and policies at the school district level do not reflect the national standards and best practices identified in research. Indeed, even into the next century, issues surrounding teacher evaluation still center on the elements of purpose, procedures, validity, and the roles of the evaluators (Peterson, 2000).

Wise, Darling-Hammond, McLaughlin, and Bernstein (1984a) determined in their case study of four districts that part of the success of an evaluation system hinged on the alignment of the purpose of evaluation with the process and procedures. The two fundamental purposes of evaluation are to improve instruction (formative) and to judge the merit and worth of a teacher (summative). Most of the research supports the need for both purposes to be addressed thus supporting the conclusions of the Wise et al., (1984a) study and the conclusions of the Joint Committee on Standards for Educational Evaluation (1988). The controversy in the literature surrounds the argument of whether both purposes can be met within one teacher evaluation system. Some researchers claim that formative evaluation demands different evaluators and techniques in order to be effective (Peterson, 2000; Popham, 1988b; Stiggens, 1985). Others theorize that given proper conditions and training, both purposes can be addressed within a single system

(Danielson & McGreal, 2000; Holland, 2005). In fact, no consensus even exists regarding the relative worth of summative evaluation juxtaposed against formative evaluation. As a result, various models of teacher evaluation have evolved reflecting a sliding scale of emphasis between the two purposes.

While the literature reflects diverse perspectives on summative verses formative evaluation, greater consensus exists on the procedures associated with teacher evaluation. The use of classroom observation as the sole data point for evaluation is firmly renounced. Yet, this practice continues to be widely practiced (Loup et al., 1996). While it is widely accepted that multiple data sources should be used in teacher evaluation procedures, the specific data sources are also an area of controversy. The literature generally agrees that the timing and intensity of the evaluations should differ given the experience and skills of specific teachers. The most contentious debate centers on the use of test results as a determinant of teacher effectiveness. Earlier studies questioned the validity of the data citing the multitude of student variables that are outside of a teacher's control (Medley & Coker, 1987). Recent studies dispute the earlier findings citing a correlation between principals' subjective evaluations and achievement test data based on a growth analysis model (Jacob & Lefgren, 2005). Most literature supports multiple, variable sources with teachers involved in determining the specific sources to be used in their evaluation (Peterson, 2000).

In order for an evaluation system to be supported by teachers and be legally defensible, the results of an evaluation must be reliable and viable. That is, it must consistently and accurately measure degrees of competence (Wise et al., 1984b). Often inconsistency occurs across schools within the same district (Wise et al.) causing distrust

and apathy towards the system. In some cases the evaluator comes under question. Systems that rely heavily on a principal's judgment have been especially scrutinized (Medley & Coker, 1988; Peterson, 2000) citing that the predispositions of a principal lead to different ratings for similar teacher practices (Wise et al., 1984b). The lack of validity in principals' ratings has been supported through studies which documented the inflation of teacher ratings (Frase & Streshly, 1994).

In addition to the challenges of maintaining consistency between evaluators, the actual level of accuracy needed in an evaluation varies depending on the utility or use of the evaluation. When applied to personnel decisions, the need for a high level of accuracy demands that evaluation criteria be standardized and applied consistently. Accuracy levels will vary depending on the criteria for receiving performance pay. If the criteria for receiving performance pay are minimal with the vast majority of teachers meeting the criteria, reliability, and validity are not as crucial. However, a high standard for awarding performance pay demands high levels of reliability and validity (Joint Committee, 1988). Finally, validity and reliability are dependent on the choice of criteria used, another area of discussion within the literature. Thus, theories regarding evaluation personnel, accuracy demands, and evaluation criteria have contributed to the development of various models and studies within the evaluation field.

Personnel Evaluation Standards

Standards have long played a role in accountability by defining high quality instruction and student learning. Reports such as the landmark *A Nation at Risk: The Imperative for Educational Reform* (National Commission on Excellence in Education, 1983) brought a wave of accountability measures into educational reform. While the No

Child Left Behind (NCLB) legislation (2001) spurred the development of academic standards for students, part of the legislation also addressed the requirement for "highly qualified" teachers. This legislation underscored the key role of the teacher in the academic achievement of students. The pressure generated by NCLB accountability measures has re-ignited schools to hire and retain highly competent teachers. In order to accomplish this goal an effective personnel evaluation system must be in place. Even prior to the NCLB legislation, there was widespread concern regarding the quality of teacher evaluation (Medley & Coke, 1983). Research supported by the RAND Corporation also supported this conclusion (Wise et al., 1984b). The Joint Committee on Standards for Evaluating Educators, comprised of nationally recognized scholars in evaluation, concluded there was a lack of agreement regarding the attributes of quality personnel evaluation. Thus, the Joint Committee undertook the challenge of developing standards in order to define the expectations of a quality personnel evaluation system. Initially issued in 1988 and revised in 2009, the personnel evaluation standards define the attributes of an evaluation system and assist users in developing and implementing systems that are reliable and valid.

The standards are designed to play a critical role in the development and implementation of personnel evaluation systems. The worth and merit of the standards results from the meticulous process applied to their development and the subsequent validation of their content. In this study, the background of the standards project will be examined first, followed by an examination of the general attributes of the standards, the content of the standards, and the application of the standards to practice and research.

Background of the Project

The Joint Committee on Standards for Educational Evaluation was created in 1975 with 16 members representing 14 professional societies. The purpose of the committee was to establish nationally recognized standards for evaluation in the following key areas: students, programs, and personnel. The Joint Committee established evaluation standards for programs and students first, addressing the personnel standards last. The sequencing of standards was intentional given the initial hesitancy of testing experts to create standards of evaluation for teachers and principals (Stufflebeam & Brethower, 1987). During the development of the first set of standards, the program standards, the Joint Committee established the definition of a standard as a principle "commonly agreed to by people engaged in the professional practice of evaluation for the measurement of the value or the quality of an evaluation" (Joint Committee, 1981, p. 12). The standards did not specify specific models or forms, but rather outlined the expectations of a quality evaluation system regardless of purpose or procedures.

The Joint Committee also defined a standard setting process which was open, public, and participatory. This process followed all the reporting and validation procedures required by the American National Standards Institute (ANSI). The seven step process included: (a) identification of the issues, (b) development of a first draft, (c) national and international reviews, (d) field trials, (e) national public hearings, (f) ANSI review, and (g) finalization of the standards (Howard & Sanders, in Stronge, 2006). An independent panel of educators oversaw the process validating that all procedures had been followed. The Joint Committee seeks input on a continual basis and reviews the standards every three years. The end result of these procedures is a set of best practices,

warnings of common errors encountered, and guidelines for the use of each standard (Howard & Sanders, in Stronge).

General Attributes and Content of the Standards

The Joint Committee established the proposition that all evaluation should have four basic attributes: *propriety*, *utility*, *feasibility*, and *accuracy*. The 27 standards are categorized by the four attributes and determine the extent to which an evaluation system meets the attribute. While any one standard could be associated with more than one attribute, it is linked to the one attribute that serves as its primary emphasis for sound evaluation. The second edition of *The Personnel Evaluation Standards* (Joint Standards, 2009) addresses the issue of diversity through the lens of all the standards. Each standard contains a statement of the standard, explanation, rationale, guidelines to help users meet the standard, a list of common errors associated with the standard, and an illustrative case of application of the standard. In addition, the second edition contains a functional table of contents which provides the user with the most common standards associated with eight applications of the standards: (a) training for evaluators, (b) certification/licensure, (c) defining roles within an evaluation system, (d) selection of an existing system of personnel evaluation by an institution, (e) development of a personnel evaluation system, (f) using evaluation results for staff development, merit awards, tenure or promotion decisions, (g) evaluating evaluates from diverse backgrounds, and (h) termination decisions. The standards can be clarified through an examination of the four attributes under which they are grouped.

The *propriety* standards address legal and ethical issues within personnel evaluations insuring the rights of the evaluatee and of those conducting the evaluation.

The seven standards grouped within the *propriety* standards address such issues as confidentiality, access to information, and comprehensiveness of the evaluation. The propriety standards also seek to assure that evaluations promote sound education for students and evaluatees are treated with respect and dignity. The seven *propriety* standards include: Standard P1 (Service Orientation), Standard P2 (Appropriate Policies and Procedures, Standard P3 (Access to Evaluation Information), Standard P4 (Interactions with Evaluatees), Standard P5 (Comprehensive Evaluation), Standard P6 (Conflict of Interest), and Standard P7 (Legal Viability).

The *utility* standards deal with the timeliness, usefulness, and weight of personnel evaluations. The primary intent of personnel evaluations is to improve educational practices for students. Thus, through these standards evaluatees are informed of practices to improve their performance. These standards also require that the purpose of the evaluation is clear whether it be personnel decisions such as tenure or professional development directions to improve practice. The standards also require that the evaluation is conducted by credible persons with appropriate expertise. The six *utility* standards include: Standard U1 (Constructive Orientation), Standard U2 (Defined Uses), Standard U3 (Evaluator Qualifications), Standard U4 (Explicit Criteria), Standard U5 (Functional Reporting) and Standard U6 (Follow Up and Professional Development).

The *feasibility* standards recognize that educational institutions have limited resources and exist in a climate influenced by political factors. The feasibility standards recognize that efficiency, funding, and political viability can impinge on the quality of a personnel evaluation systems. The three *feasibility* standards include: Standard F1

(Practical Procedures), Standard F2 (Political Viability), and Standard F3 (Fiscal Viability).

The purpose of the *accuracy* standards is to insure that evaluations result in a valid judgment of the instructional performance of the evaluatee. The evaluation should be legally defensible with conclusions based upon relevant data. Furthermore, the evaluation should fit its stated purpose and be free of bias. The eleven *accuracy* standards include: Standard A1 (Valid Judgments), Standard A2 (Defined Expectations), Standard A3 (Analysis of Context), Standard A4 (Documented Purposes and Procedures), Standard A5 (Defensible Information), Standard A6 (Reliable Information), Standard A7 (Systematic Data Control), Standard A8 (Bias Identification and Management), Standard A9 (Analysis of Information), Standard A10 (Justified Conclusions), and Standard A11 (Metaevaluation).

Application of the Standards

The personnel evaluation standards serve multiple needs for multiple audiences. The applicability of the standards falls generally into one of two categories, development of evaluation systems or review of evaluation systems. In addition to these uses *The Personnel Evaluation Standards* 2nd Edition also cites a research application for the personnel evaluation standards. The standards provide a framework for structuring logical arguments and hypothesis for studies that examine teacher evaluation. Indeed, Ellet, Wren, Callender, Loup, and Liu (1996) used the standards to examine the System for Teaching and Learning Assessment and Review (STAR) that was developed and implemented in Louisiana. In 1997 Loup and Ellett completed a similar study of 14 case studies analyzing the applicability of the standards in Connecticut school districts.

Included in *The Personnel Evaluation Standards* 2nd Edition is a metaevaluation case study conducted in 2005 which applied the standards to the procedures, practices, and related components of a school district in order to determine the soundness of the personnel evaluation system. The study also investigated the perceptions of the district personnel concerning the evaluation system. These studies are significant in establishing the personnel evaluation standards as credible criteria for framing research in the area of teacher evaluation.

Approaches to Evaluation - Models

Purpose and philosophical underpinnings drive various approaches to evaluation. Criteria, process, forms and instrumentation, and evaluative personnel are components that differ across assorted studies and theoretical literature. The literature also reflects a wide assortment of constructs in categorizing the various approaches to evaluation. For example, Wheeler and Scriven (in Stronge, 2006) cite eight philosophical foundations upon which upon which evaluations systems are categorized. In contrast, Scriven's (1995) 12 approaches to evaluating merit categorizes models by types of criteria and evaluator roles. Models are sometimes simply defined by one criterion usually based on whether the model focuses on what teachers are doing or the results they are obtaining (Danielson & McGreal, 2000). In this study, systems of evaluation will be categorized and analyzed based upon the criteria and process applied within the system.

Criteria-based Approaches

The evaluative criteria applied to a teacher evaluation system are directly related to a philosophy or a stated purpose of evaluation. Criteria used for judging minimal competency must be generalizable, standardized, and legally defensible. Criteria used to

meet improvement objectives must be subject area and grade level specific (Wise et al., 1984). The source of the criteria can also range from self identified goals to specific duties defined by the organization. Research has shown that teacher support for evaluation systems diminishes when the criteria does not fit in the organizational context (Loup et al., 1996; Wise et al.). Thus, examination of the criteria provides a way to analyze and compare evaluation approaches. For purposes of this study, criteria will be further divided into performance-based and results-based criteria.

Various organizations have worked to formalize the definition and components of effective teaching. In several cases these efforts began with the establishment of standards for initial licensure. In 1987, the Educational Testing Service (ETS) drew on effective teaching research to develop the Praxis Series: Professional Assessments for Beginning Teachers. The National Board of Professional Teaching Standards (NBPTS), established in 1987, was one of the first organizations to develop a standards-based approach to identify quality teachers. The standards underwent rigorous review and a comment period prior to their adoption. The purpose of the standards was not to replace state licensing requirements, but rather to establish advanced standards for experienced teachers. Teachers seeking certification as a National Board Certified Teacher undergo a rigorous certification process that includes the development of a portfolio that demonstrates their classroom decision-making processes and instructional skills. The number of National Board Certified Teachers has substantially grown (over 10,000 in 2003) with many school districts offering additional compensation. While the standards used for the National Board Certification have been incorporated into other evaluation instruments (Danielson, 1996; Stronge & Tucker, 2003), little research is available to

validate the effectiveness of this process in identifying effective teachers. A limited study of six National Board Certified Teachers (Pool, Ellett, Schiavone, & Carey-Lewis, 2001) showed considerable variation in their skills and practice. Results from this study indicate that two teachers were exemplary, two were average, and two were deficient in some areas. Clearly, additional studies need to be conducted in order to further determine the linkage between the certification and quality teaching.

Performance-based models compare a teacher's actions and knowledge to a set of descriptors. Evaluators judge a teacher's merit against commonly accepted criteria. Standards-based evaluation is one form of a performance-based model. Based on effective teaching research, standards-based evaluation defines the criteria based on research which correlates teacher traits with high student achievement levels. Danielson's (1996) *Enhancing Professional Practice: A Framework for Teaching*, which is based on The Interstate New Teacher Assessment and Support Consortium (INTASC) is a widely known and researched example. Stronge and Tucker's (2003) *Handbook on Teacher Evaluation: Assessing and Improving Performance* also frames an evaluation system based upon domains, standards, and performance indicators. Stronge and Tucker's work organizes the tools and rubrics from the perspective of a job description for teachers. Research is used to support and inform the performance standards. Opponents of standards-based models argue that teachers must adjust to the context in which they teach. In addition, researchers note that correlation studies yield knowledge but not certainty as to the specific practices of a teacher that result in higher achievement. A 2004 study (Kimball, White, Milanowski, & Borman, 2004) analyzed the relationship between scores on a standard-based evaluation system modeled on the *Framework for Teaching*

(Danielson, 1996) and student achievement levels. The results of the study were mixed as to the relationship between teacher evaluation scores and the average achievement of those teacher's students. A positive association existed between the teacher evaluation scores and the average student achievement scores, but the coefficients were not statistically significant in all cases. A curricular alignment variable was cited in an elementary study of evaluation scores and student achievement (Gallagher, 2004). In this study, tightly defined curriculum and test alignment contributed to a finding of statistical significance between teacher evaluation scores and achievement levels. Citing classroom context effects, Borman and Kimball (2005) concluded that standards-based approaches may skew the ratings of teachers in less advantaged classrooms because of the attributes of the students they are teaching. Studies of the implementation of standards-based systems are not conclusive. While teachers generally did appreciate the comprehensiveness and specificity of competencies (Kimball, 2002) other teachers' reactions were neutral or unfavorable (Heneman & Milanowski, 2003).

Citing concerns regarding evaluation systems based on effective teaching research, Scriven (1994) proposed the use of a teacher evaluation approach based upon the duties of a teacher. This approach, the Duties Based Teacher Evaluation model (DBTE) based upon a list of duties of the teacher (DOTT), was developed in the 1980's and refined in the 1990's. Proponents of DBTE refuted the idea that research can define the characteristics of a good teacher insisting instead that the research only highlights the relative importance of various styles of teaching (Scriven, 1995b). DBTE is not founded on research but rather on an explicit job description and implicit job obligations. The list of duties upon which the model was initially based was circulated and revised by

thousands of stakeholders including teachers, administrators, parents, lawyers, and students in Australia and the United States. The work was later revised again in light of the PRAXIS project and the Teacher Evaluation Models Project, a part of the work at the Center for Research on Educational Accountability and Teacher Evaluation (CREATE). Once again, the resulting list of duties was not a job analysis but rather a list of what teachers can legitimately be held responsible for knowing and being able to do. The DOTT is organized into five general areas of competence: 1) knowledge of subject matter, 2) instructional competence, 3) assessment competence, 4) professionalism, and 5) other duties to the school and community. These general areas are further divided into sub areas and sub elements. The philosophical underpinning of this model asserts that teachers can teach using any style that is ethical and effective given the context of their setting. While DBTE is frequently cited in the literature, no studies regarding the feasibility or utility of this approach were located. The significant role of the work rests on its contributions to the PRAXIS project and the National Board of Professional Teaching Standards.

While teacher quality can be conceptualized in terms of what teachers do, in other words, their behaviors or performances, teacher quality can also be conceptualized in terms of the results that they produce. This concept has been similarly termed “teacher outcomes” (Stronge, in Stronge, 2006), “results-orientated” (Heneman, 2003), and “objective-measures” (Bommer, 1995). Studies have consistently affirmed that teachers are a key factor influencing achievement levels within a classroom (Nye, Konstantopoulos, & Hedges, 2004; Wright, Horn, & Sanders, 1997). Thus, while teachers historically have not considered standardized test scores to be a valid measure of teacher

effectiveness (Wise et al., 1984b), a number of school districts began using results as one component of their teacher evaluation system in the 1990's (Stronge & Tucker, 2005). Measurements of teacher quality have shifted from looking solely at a teacher's characteristics and performance to a teacher's impact on student learning as public pressure for accountability has increased.

Results-based evaluation criteria use data on student progress toward mastering instructional goals and objectives. Evaluation system utilizing results-based criteria are not generally supported by teachers (Stronge & Tucker, in Stronge, 2006). Problems cited include the lack of valid student gain data, influences on student achievement outside of teacher control, and the minimization of other factors or dynamics that are equally as important in the teaching and learning process (Peterson, 2000). For purposes of this study, results-based criteria literature will be reviewed in terms of the personnel evaluation standards (Joint Standards, 2009). Analysis will be framed within the standard's four attributes of *propriety, utility, feasibility* and *accuracy*.

Studies reveal that the main concerns regarding the use of achievement tests as a result-based criteria are found across all four standard attributes. The greatest number of concerns falls within the accuracy standards specifically around Standard A5 (Defensible Information), Standard A9 (Analysis of Information), and Standard A10 (Justified Conclusions). Several studies have provided evidence supporting the accuracy of using achievement test data as an evaluative criterion. The Tennessee Value-Added Assessment System (TVAAS) generated a series of supportive studies. The TVAAS developed by William Sanders used a statistical model which yielded scores based on growth or gains scores rather than on a fixed standard. The TVAAS database recorded each individual

student's growth from second to eighth grade. An initial study (Sanders & Rivers, 1996) suggested that the teacher effects are both additive and cumulative. The evidence suggested that residual effects of both effective and ineffective teachers were measurable two years later. In a study (Wright, Horn, & Sanders, 1997) using the TVAAS, the evidence indicated that race, socioeconomic level, class size, and classroom heterogeneity are poor predictors of student academic growth. The paramount finding was that teacher effectiveness was once again the major influence on student gain. Finally, a third study (Sanders & Horn, 1998) affirmed the influence of teacher effectiveness by analyzing cumulative gains for schools across the entire state with regards to racial composition, percentage of students receiving free and reduced lunches, or the mean achievement level of the school. Based upon the evidence all three studies recommended the inclusion of student achievement data in teacher evaluations given the data are generated through a standardized testing program, administered longitudinally, and appropriately analyzed. In order to further insure that effect differences were not a result of differences in students, Nye, Konstantopoulos, and Hedges (2004) randomized the assignment of students and teachers to classrooms. Evidence from the study further supported the position that student achievement data can serve as an indicator of the substantial differences in the effectiveness of teachers.

The integrity of value-added modeling (VAM) systems such as TVAAS was challenged through evidence produced in a meta-analysis (McCaffrey, Lockwood, Koretz, & Hamilton, 2003) sponsored by the RAND corporation. The six VAM studies involved in the meta-analysis included the TVAAS studies. VAM systems are generally described as using complex statistical techniques over longitudinal achievement data in

order to determine teacher effects. The evidence did not support the conclusions reached in the six VAM studies. Based upon what the researchers determined to be faulty statistical techniques, the study concluded that VAM-based rankings of teachers are highly unstable and should be used with used with extreme caution. While the researches did not support the use of VAM for high-stakes decisions, they did conclude that VAM estimates could be used for lower stakes decisions such as use as a preliminary filter for identifying teachers needing a more thorough review. With regards to teacher influence on student achievement, the study concluded that there was little convincing evidence of the magnitude of teacher effect. Thus, the research is inconclusive regarding the accuracy of VAM data in determining teacher effectiveness. The contradicting evidence appears to stem from disagreements regarding the validity of the statistical techniques. The same studies also question the ability of local districts or even states to develop sophisticated testing systems that would accurately reflect the district's or state's curriculum.

Studies also indicate problematic areas in both the *propriety* (Standard P7- Legal Viability) and *utility* (Standard U4- Explicit Criteria) standards. The standards identify stakeholder agreement as an aspect of Standard P7 (Legal Viability). While the Ohio Federation of Teachers and the Ohio Education Association in 2004 did embrace a VAM system as one component of the teachers' evaluation system, teachers are generally reluctant to embrace student achievement data as an evaluative criterion (Peterson, 2000; Stronge, in Stronge, 2006). Standard U4 (Explicit Criteria) requires that the criteria for one group may not be appropriate for another group. For example, teacher criteria should not be applied to counselors. In applying this standard, it would be difficult for all teachers to have the same criteria as not all teachers are responsible for the tested areas.

In spite of the arguments on both sides, achievement data continues to attract a great deal of attention from policymakers and legislators.

In a follow up study on the state of teacher evaluation in the 100 largest school districts Loup, Garland, Ellet, and Rugutt (1996) note that while parents potentially are a valuable data source in teacher evaluations, districts have not systematically incorporated parent surveys in evaluation systems. Few studies within the past 15 years have addressed the use of parent survey data in evaluation systems. The literature generally connects parent data with three *accuracy* standards and one *propriety* standard: Standard A5 (Defensible Information), Standard A6 (Reliable Information), Standard A8 (Bias Identification and Management), and Standard P7 (Legal Viability). In the research conducted on parent surveys teachers chose to use parent survey data (Ostrander, 1996; Peterson, 2005). Teachers were pleased with the results and surprised at the quality of the information (Peterson, 2005). In an attempt to establish the accuracy of parent ratings parent, student, teacher self-rating, and principal ratings were compared (Ostrander, 1996). Students tended to score teachers the lowest followed by parents. Teachers and principals gave the higher ratings with principals giving the highest ratings. Peterson (2005) further cautions the use of parent data as a sole evaluative criterion citing the declining return rate at higher grade levels. The decrease in contact and communication at the secondary level results in more global or halo ratings. While parent survey data gives insight into teacher communication and the parent perspective, there is insufficient research to support its use other than as one of multiple sources of evaluative data.

Another source of results-based criteria is student survey data. In a study of the teacher evaluation in the 100 largest school districts (Loup et al., 1996), less than 5% of

the school districts systematically used student views as part of teacher evaluation. Yet, some researchers ascertain that, based upon student achievement data, student ratings constitute more accurate feedback than that from others (Wilkerson, Manatt, Rogers, & Maughan, 2000). Research points to two attributes within the standards for consideration in the use of student surveys for results-based criteria: *accuracy* and *feasibility*. Within the *accuracy* attribute, it is widely agreed that there exists a correlation between parent and student ratings while noting the lack of correlation with principal and teacher self-rating (Peterson, 1987; Ostrander, 1996; Wilkerson et al.). Most evidence suggested that students rated teachers consistently lower than principals. However, Peterson (1987) reported higher student ratings than principal ratings. Reasons cited for the lower student ratings include the unique position of students as clients and benefactors of the teachers' work. Correlation levels between two years of surveys suggest that more than two years of data are needed to establish stable patterns of results (Peterson, Wahlquist, & Bone, 2000). While correlation or lack thereof between raters does not necessarily justify accuracy, Wilkerson et al. extended the research to include comparison of rater results to performance of K-12 students on criterion-referenced reading, language arts, and mathematic tests. The evidence from this study found student ratings of teachers to be the best predictors of student achievement. Furthermore, the evidence revealed student ratings have the strongest positive relationship to student achievement when compared to the ratings of principals and self-ratings of teachers. With regards to Standard A5 (Defensible Information) and Standard A8 (Bias Identification and Management), limited research indicates that the surveys were not popularity contests but rather recognition of who enables a student's learning (Peterson, 2000). The criteria for positive ratings did

shift between secondary students and elementary with older students placing higher value on learning and younger students placing higher value on relationships (Peterson et al.). Ultimately, the use of student data is driven by politics and contextual elements (Peterson et al.). In all the studies teachers chose to use student survey data for evaluative feedback. It is unknown if Standard F2 (Political Viability) could be met if student survey data were required systematically.

Results-based criteria by definition are a factor of outcomes teachers produce as a result of their work. Some criteria, based upon their use or composition, could be considered either as a part of the process of teacher evaluation or as a criteria for teacher evaluation. Teacher portfolios fall into this category. Tucker and Stronge (in Stronge, 2006) categorize portfolio assessments as Teacher Outcomes while categorizing portfolios as Teacher Behaviors. Indeed, Peterson (2000) defines a teaching portfolio designed for evaluative purposes as containing teacher selected materials, products, artifacts, and reflections on work. Portfolios may also include multiple data sources. Student and teacher work is also included in a descriptor crafted by Wolf (in Stronge, 2006). Thus, if portfolios consist of objective results-based data such as student surveys and achievement results, they could be categorized as being results-based. Research studies typically fail to include the specific guidelines for portfolio development. In fact, numerous studies have cited the lack of direction and clarity in portfolio development as a disadvantage of portfolios for teacher evaluation (see, for example, Heneman & Milanowski, 2003; Kimball, 2002). Numerous studies also cited lack of clarity of purpose and intense workload for minimal value (Heneman & Milanowski, 2003; Kimball, 2002). Research from a Florida study on the implementation of a teacher evaluation system

(Davis, Pool, & Mits-Cash, 2000) noted some participants labeled portfolios as "worthless wastes of time." While this review does not serve as an evaluation of portfolios, the research clearly identifies issues regarding lack of focus, lack of benefit, and lack of procedural clarity in portfolio use. These elements are addressed in the standards specifically in the areas of *utility*, *feasibility*, and *accuracy*.

A related, yet different approach to data collection takes the form of a dossier organization technique. Citing and experiencing the widespread challenges in the use of portfolios in Utah, Peterson (2001) explored the use of the dossier approach. A dossier is a collection of documents assembled with the specific purpose of attesting to teacher quality. Data and information is collected specifically to make value judgments regarding teacher quality. Dossiers are more compact and focused. While the dossier approach may be suited for use as results-based criteria, the dearth of research in this area prevents any substantiation of this theory.

Process-based Approaches

The philosophical underpinnings of an evaluation system dictate not only the criteria used to determine the value and merit of a teacher but also the process by which that determination is made. Historically, the determination was made using one universally applied teacher evaluation process in which a single administrator observed a teacher in a classroom setting. The trend in teacher evaluation is away from this practice. Two major themes have emerged that reflect a change in this historical process: (a) the use of multiple and variable lines of evidence and (b) differentiation in evaluation procedures with regards to focus, intensity, and selection of the evaluator. School districts and state level policymakers have moved to align the policies and procedures of

evaluation systems to institutional missions in response to the accountability movement, a renewed focus on professionalism, and research on factors influencing instructional improvement.

In a follow-up study of teacher evaluation systems in the 100 largest school districts, Loup et al., (1996) concluded that the methods of teacher evaluation still relied heavily on direct observation of teaching and informal observations of teachers. Results of surveys indicated that most districts used both direct observation of teachers (94%) and informal observations (86.8%) as part of their evaluation process. Other frequently cited processes included teacher self-rating (45.6%), student achievement data (25%), and teacher portfolio assessments (23.5%). Few districts reported using multiple lines of evidence other than formal and informal classroom observations. This practice runs contrary to the widespread agreement and theoretical positions of numerous researchers. Indeed, the evidence from numerous studies supports the use of multiple data sources asserting that various lines of evidence assess different concepts of quality. A premier study in this area (Peterson, 1988) examined the results of six lines of evidence including administrative reports, pupil surveys, parent surveys, teacher tests (National Teachers Exam), professional activities, and years of experience. The mean absolute correlation between the lines of evidence was .15 suggesting that the data examined different dimensions of teacher quality. Further support of this proposition was provided by studies of student survey data (Peterson, Wahlquist, Brown, & Mukhopadhyay, 2003), and parent survey data (Peterson & Peterson, 2005). In the student survey data, evidence showed that younger students rated teachers with a focus on the quality of relationships established with students while older students weighted their ratings on teachers' ability

to deliver clear instruction (Peterson et al., 2000). In contrast, parent feedback focused on three important concerns: humane treatment for students, support for student learning, and effective communication and collaboration with parents. Correlations between student achievement data and the evaluator's rating of teachers ranged from 0.30 to 0.40 in two studies- one conducted in Cincinnati (Milanowski, 2004) and the other in a small charter school (Gallagher, 2004). A third study (Kimball et al., 2004) provided only tentative evidence of a correlation between the evaluator's ratings and student achievement. The higher correlations realized in the small charter school were partially attributed to the tight alignment between evaluation criteria, curriculum, and assessments. Evidence suggested that the lower correlations in the Kimball et al. (2004) study were partially attributable to the lack of such an alignment. In sum, these studies support the theory put forth by Peterson (2000) and Scriven (1994) that multiple lines of evidence reveal different aspects of teacher quality thus providing a clearer and more valid teacher evaluation process.

A prominent study of effective practices in teacher evaluation (Wise et al., 1984) set the standards for multiple change efforts in evaluation systems. Based on the evidence, the study recommended evaluation procedures be closely aligned with the purpose of evaluation whether that be for instructional improvement or judgments regarding teacher quality. The study further questioned whether one system could meet the needs of all teachers. Evaluations for teachers who are competent require evaluators who have subject and grade level expertise. Principals or other evaluators with generic teaching knowledge are less able to differentiate degrees of expertise within this group of teachers. The evaluative criteria for teachers at risk of termination must reflect the

minimally acceptable teaching behavior and be legally defensible. The study concluded that evaluations must be flexible in order to take into account the specific teaching context, move teachers forward, and meet the specific needs of the district. Thus, subsequent research in teacher evaluation has addressed variations in focus, intensity, and evaluators.

The focus of the evaluation should determine many of the activities that are included in the process. Most evaluation systems differentiate between tenured and non-tenured teacher evaluation. Frameworks proposed by Danielson (2000) and Stronge and Tucker (2003) both differentiate between beginning and experienced teachers. Danielson further defines a track for teacher assistance. Both frameworks identify the primary purpose of the beginning teacher evaluation process as a means of obtaining usable and reliable data to assist in making tenure or continuing contract decisions. The focus of the tenured teacher evaluation is to assist competent teachers in becoming even more competent. The number and nature of the evaluative activities differ based on the focus. Field studies of teacher evaluation systems also commonly note a greater number of evaluations required of beginning teachers verse their tenured counterparts. However, besides a variation in frequency of evaluations, only a few studies document differences in processes within the two categories (Danielson; Stronge & Tucker). Differentiation can also occur through allowing teachers choices. Several studies examined the use of variable lines of evidence in which teachers choose the data to be used in their evaluations (Kimball, 2000; Peterson, 1987) or select the area in which they were to be evaluated (Henneman & Milanwski, 2003).

Perhaps the most controversial area of differentiation in teacher evaluation rests with the evaluator. While it is generally agreed that principals' expertise is adequate to assist beginning teachers, evidence also indicates that evaluators with subject area expertise are needed to further promote the growth of competent teachers (Wise et al., 1984). In spite of the research, principals still serve as the primary evaluators of teachers (Wilkerson et al., 2000). However, research has also revealed that even if the evaluator is a teacher and has subject area or grade level expertise, teachers question the validity and reliability of the evaluator's ratings (Heneman & Milanowski, 2003). The validity of principal ratings has also been questioned by teachers. This challenge is supported by evidence that principals consistently rate teachers higher than teachers' self-rating, students, or parents (Peterson et al., 2003; Peterson & Peterson, 2005; Wilkerson et al.). Studies have speculated that the conflict of leadership roles and evaluative roles may be responsible for this variance (Wise et al.). The controversial nature of the ratings of evaluators is underscored by the fact that in three separate extensive studies of the implementation of a standards-based evaluation system, the validity of evaluators was consistently an issue regardless of their role or knowledge base (Davis, Ellett & Annunziata, 2002; Heneman & Milanowski, 2003; Milanowski & Heneman, 2001). The tenacity of this issue supports the conclusions of Ellett, Wren, Callender, Loup, and Liu (1996) that the climate and educational context creates perceptions that data cannot overcome.

Principals and Teacher Evaluation

Principals have played a significant role in the evaluation of teachers for the past 20 years (Colby, Bradshaw, & Joyner, 2002). Historically, the principal played a

managerial role within teacher evaluation systems. The effectiveness of principals' evaluations depends on whether they consider evaluations a burden or an opportunity for professional growth (Davis et al., 2000; Wise et al., 1984). Recently, the role of the principal has become one of an instructional leader rather than a manager. This shift has resulted in a role conflict with principals acting as an advisor for improved instruction while at the same time serving as the judge of competence. The resulting political and sociological factors of the role conflict have contributed to the teachers' perception that principals are ineffective evaluators due to lack of sufficient resolve and competence to evaluate accurately (Wise et al., 1984). Yet, the importance of the principal's leadership in the successful implementation of a teacher evaluation system is not disputed by the literature. In particular, a principal's attitude and approach towards a system can make the difference between a perfunctory performance assessment system and a meaningful assessment system that has the potential to improve teaching and enhance learning (Davis et al., 2002). With the recent adoption of state accountability systems that reward and sanction schools and provide incentive compensation for teachers, the role of the principal has come under question (Jacob & Lefgren, 2008). While unions support peer coaching and peer review as an evaluative process for determining teacher effectiveness (Wilkerson et al., 2000), studies addressing the role of the principal in teacher evaluation can inform principal's roles in compensation, personnel decisions, and instructional leadership. The success of an evaluation system utilizing principals as evaluators largely depends upon the ability and willingness of a principal to identify the quality of teachers.

This section examines literature which addresses the factors relating to the challenges of and supports for principal conducted teacher evaluations. Issues are

examined in light of the technical aspects and sociological aspects of principal conducted teacher evaluations. The section concludes with issues specific to beginning and marginal teachers.

Technical Aspects

Research over the past 20 years sought to sort out the technical capacity of principals to judge teachers. The adequacy of principals in terms of expertise, the processes applied, the actions taken or not taken have all come under scrutiny. In order establish a sense of accuracy with principal conducted teacher evaluations studies have examined principals' ratings in comparison to student achievement scores, experience and compensation levels, and client ratings. Additionally, evidence supports possible biases in judgment based upon values, gender, years of experience, and the influence of politics. The following sections explore these issues.

In an attempt to validate the accuracy of principal conducted evaluations several studies have correlated teacher ratings with student achievement data. These studies are based upon the assumption that quality instruction yields increased student achievement as measured through student achievement tests. In this case, an effective teacher is defined as one that can produce high student achievement scores. The results of these studies are inconclusive due to variation in tests, statistical methodology, and alignment between the curriculum and the assessment. In an earlier study Medley and Coker (1987) found that teacher's ratings and standardized test results were correlated 0.20 with no significant differences in the rating accuracy of the 46 elementary principals in the study. The study concluded that the evidence provided no proof that the average principal was a good judge of teacher quality. A subsequent study by Manatt and Daniels (1990)

challenged this proposition by citing limitations such as contamination of data caused by compiling data across schools, lack of curriculum and test alignment, invalid assumptions, and questionable statistical procedures in the Medley and Coker study. Manatt and Daniels accommodated for these limitations by utilizing evaluator training to negate interschool difference and criterion reference pre and post test data. As opposed to the Medley and Coker's study, Manatt and Daniels concluded that principals can accurately evaluate the performance of teachers given extensive training and criterion test data.

Approaches to testing took a significant shift with the application of value-added measurement systems (VAM) employing statistical advances as applied in the study of the Tennessee Value-Added Assessment Database (Sanders & Horn, 1998). Successive studies correlating gain results from value-added assessment systems to principal-based teacher ratings generally supported the proposition that rating scores can be substantially related to student achievement gains (Gallagher, 2004; Jacob & Lefgren, 2008; Milanowski, 2004). Moreover, the variations in teacher effects were reflected more through the principal-based ratings than through teacher education and experience (Kimball et al., 2004). Jacob and Lefgren (2005) further expanded these conclusions to examine the predictability of future student achievement based upon principals' ratings, teacher experience, education or actual compensation, and value-added teacher quality measures. While value-added teacher quality measures proved to be the best predicator, principal-based ratings were a significantly better predictor than teacher experience, education, or actual compensation. Furthermore, principals were good at identifying those teachers producing the smallest and largest achievement gains (top and bottom 10-

20 percent) while less able to distinguish between the middle 60-80 percent. Correlations between principal-based ratings and value-added measures were between 0.29 and 0.32 for reading and math (Jacob & Lefgren, 2008). In contrast, a study of the relationship between student achievement on criterion-referenced tests and performance ratings by principals, students, and teacher self-rating indicated that the best predictor of student achievement on district-developed, criterion-referenced tests was student ratings (Wilkerson et al., 2000). Moreover, student ratings also showed the strongest positive relationship to student achievement when compared to teachers and principals.

The literature reflects a lack of widespread agreement on the level of correlation needed to substantiate the relationship between principal-based ratings and teacher effectiveness. For example, while correlations of .29 and .32 (Jacob & Lefgren, 2008) were cited as showing a strong relationship, Heneman (1987), in a classic meta-analysis of 23 studies of relationships between supervisory ratings and results-orientated measures, concluded that a correlation of .27 was not strong. A subsequent meta-analysis in 1995 (Bommer, Johnson, Rich, Podsakoff, & MacKenzie) of studies containing both objective and subjective ratings of employee performance resulted in a corrected mean correlation of .389. Again, the Bommer et al. study concluded that the value indicated that objective and subjective performance measures should not be used interchangeably. These studies conflict with the correlation assumptions of the value-added teacher studies.

The personnel evaluation standards (Joint Committee, 2009) contain 11 standards addressing issues of *accuracy* in personnel evaluation. Standard A8 (Bias Identification and Management) specifies that the results of evaluations should not be influenced by

preconceived ideas nor should they be based on information unrelated to the actual job performance of a teacher. However, the extant literature suggests that some bias does exist with regards to principal-based evaluations. Indeed, over 20 years ago Scriven (1981) cited the personal biases of the evaluator as one the main problems with accuracy in personnel evaluation.

There are a limited number of studies addressing the biases of principals as evaluators. Research has generally addressed areas of gender, tenure status, and principal experience. There is a lack of widespread agreement on the nature of gender bias in principal conducted evaluations. Some evidence suggests that principals systematically discriminate against male teachers (Jacob & Lefgren, 2005; Rinehart & Young, 1996). Some evidence indicates that female evaluators rate both males and females significantly lower (Manatt, 1988). Yet, Jacob and Lefgren (2005) found that both male and female principals rate male teachers lower but only on their ability to raise student test scores, not on the overall rating. Tenure status studies are more conclusive with principals systematically discriminating against untenured faculty (Bridges, 1992; Jacob & Lefgren, 2005). Yet, evidence suggests that neither the years of experience nor the duration of time the principal has known the teacher affects ratings (Jacob & Lefgren, 2008). The experience of the principal does seem to affect their overall ratings. Evaluators in the 11-15 year cohort are tougher than their colleagues with fewer years (Manatt, 1988). Moreover, Jacob and Lefgren (2008) found male principals who have been at their schools for less than four years and are confident in their ability to assess teacher performance do appear to rate teachers more accurately.

While these findings are suggestive, they also have limitations that should be taken into consideration. Most studies have small samples and, in some cases, are non-representative samples. Most importantly, they often do not account for measurement error which can skew the conclusions of the study.

Sociological Aspects

While numerous studies have examined technical questions of principal conducted teacher evaluation, literature also points to the sociological aspects of principal-based evaluations as an obstacle to quality teacher evaluation. While principals may have the ability to identify the quality of teachers, sociological factors may influence their willingness to do so. The following section explores two facets related to sociological problems with principal conducted teacher evaluation: principal resolve and teacher perceptions.

The willingness of a principal to rate teachers accurately is affected by their roles, relationship with staff, and level of authority. This assertion is supported by research conducted by Medley and Coker (1987) and Ostrander (1996) which concluded that principals' ratings are generally upwardly biased. Various other studies have controlled for this effect by stipulating in their methodology that principals' ratings were for research purposes only and were not to be applied to any sanctions or rewards for teachers (Minanowski, 2004). These studies often cited this variable as a limitation speculating that, minus the research component, principals may have rated differently given the influence of sociological affects of sanctions and rewards. Other research speculates that results may vary if evaluations of principals included their teacher rating accuracy (Jacob & Lefgren, 2008). The angst appears to stem from role conflicts and

personal characteristics. Principals as instructional leaders serve to support and encourage staff. When principals also serve as summative evaluators imparting rewards and sanctions, teachers become more guarded and conservative around the principal (Peterson, 2000). Human nature seeks to avoid conflict. Depending on the organizational culture, employee sanctions and rewards can cause a high level of discomfort that not all principals are willing to experience. Regardless of the technical quality of an evaluation system, evidence suggests a principal's attitudes and beliefs about the system and their own values and beliefs in their competency will significantly impact a principal's will to accurately identify the quality of a teacher (Painter, 2000).

Another factor that influences a principal's will to rate staff accurately is the value placed upon the task and perceived benefit of doing the task. The literature yields varying positions regarding the principal's will to engage in teacher evaluation. These positions range from regarding teacher evaluation as a waste of time (Wise et al., 1984) to registering a high level of commitment (Painter, 2000). The lack of concurrence in the studies may be due to the variation in focus. The focus of Wise et al. was on teacher evaluation in general while Painter specifically targeted evaluation to identify and take action on incompetent teachers. Within the scope of teacher evaluation, a principal has limited power. While charged with the responsibility for conducting evaluations, they rarely have a strong role in the politics of overseeing the evaluation process. The results of their evaluations typically only serve as recommendations for re-employment and dismissal. Even these decisions are limited legally, especially with regard to tenured staff. While some principals possess power to determine rewards such as additional compensation, the additional money is usually small and not a significant factor (Bridges

& Groves, 1999). Depending upon the employee contract, principals do have some power through their ability to assign positions. The main source of principals' power is generated through doing their jobs effectively, being responsive to the needs of the parents and teachers, and supporting the staff emotionally. A staff member who receives a negative rating can diminish this power by undermining the principal. Thus, principals consider possible professional consequences of negative staff ratings given current organizational support and climate.

A final factor in the sociological aspects of teacher evaluation is teachers' perceptions. If teachers perceive an evaluation system to be invalid no classical studies of validity will overcome this belief. A study of the failed Louisiana state teacher evaluation system (Ellett et al., 1996) concluded that validity contains a variety of elements. In this case the perception by those affected by the evaluation decisions was more important than the technical studies of validity. The perception of validity was a key element in the ability to maintain the viability of the system. In a similar vein, teachers have perceptions of the expertise of principals as judges of teacher quality. Teachers' perceptions of evaluator fairness permeate the literature (Heneman & Milanoski, 2003; Wise et al., 1984; Zimmerman & Deckert-Pelton, 2003). Commonly cited concerns are lack of subject area and grade level expertise, inadequate understanding of classroom context based upon the school composition, and the timing of evaluation. Even given extensive training and inter-rater reliability verification procedures, perceptions of validity will prevail given a particular political climate and context (Heneman & Milanoski, 2003).

Consensus within the Field

While a lack of widespread agreement exists in many aspects of principal conducted teacher evaluation, there are several themes of agreement within the literature. With respect to the role of the principal, studies identify the evaluation of beginning and marginal teachers as a key function of the principal. The evidence supports the effectiveness of the principals in identifying the basic skills required to be a competent teacher (Jacob & Lefgren, 2008). Furthermore, principals are more apt to initiate personnel actions based on the quality of non-tenured staff than that of tenured staff (Tucker, 1997). While studies have consistently suggested principals avoid confronting incompetent tenured teachers (Bridges, 1992; Painter, 2000), principals also are consistently noted as a having a key leadership role in addressing marginal tenured staff. The source of legitimacy of the principal's evaluation role lies in its being a legally mandated practice through state statutes (Holland & Garman, 2001).

Finally, some consensus does exist in the validity of principal conducted teacher evaluations. While studies correlating principal ratings to student achievement levels have been inconclusive, studies correlating student achievement levels to teacher experience and compensation have consistently concluded that principal ratings are a better predictor of future student achievement levels than either experience or compensation (Jacob & Lefgren, 2008; Kimball, 2004). Principals are also in a unique position to observe both the inputs and outputs of the educational process. Controversy exists as to whether principals should have a narrow evaluative focus addressing only those teachers who are beginning or marginal (Peterson, 2000) or whether they should have an expansive authority to incorporate principal evaluations into teacher promotion

and compensation systems (Jacob & Lefgren, 2008). Yet, the literature is clear – principals are expected to be instructional leaders while setting standards for high quality teacher evaluation within their schools.

Teacher Performance Pay and Teacher Evaluation

The NCLB legislations triggered a resurgence of interest in teacher performance pay. Long a standard in other professions, education has widely used a standard single salary schedule since its inception in the 1950's. Teacher performance pay is based on the theory that compensation based on skills and knowledge will attract and keep talented people in education. In addition, some theorists believe performance pay motivates current teachers to become more effective. The development of performance pay systems does have implications for principals. Some of the performance pay programs use student achievement scores or gains as one of several factors in determining teacher performance while others link teachers' salaries solely to achievement. Given the lack of consensus regarding the validity of using student test scores for performance pay, another option is to link subjective principal conducted teacher evaluations to compensation. Historically, principal evaluations have had minor impact on tenured staff in terms of rewards or sanctions. Linking performance pay to a principal's rating of a teacher would create high stakes evaluations resulting in a new level of accountability for principals in an area that previously was low impact.

Rationale for Performance Pay

Teacher compensation has evolved from the "room and board" compensation model of the early 19th century to a grade based compensation model in the late 19th to early 20th century. Both of these models were a product of the time in which they existed.

In the early 19th century the majority of citizens living in rural areas were farmers. Teachers received a small stipend and room and board by rotating their residence with the various farm families. This one-room schoolhouse model reflected the population density of the time and the need for child labor to sustain the farms. With the industrialization of the late 1800's and early 1900's came a new model of teacher compensation based upon the skills needed to provide instruction at the various levels. Because it was believed that elementary students were easier to teach and required less knowledge, secondary teachers were paid at a higher level. While this model did result in the beginning of standardization of salaries, it also resulted in gender and racial inequities as white males had more opportunity to receive the training needed to teach at the secondary level. Finally, with advent of labor unions and improved working conditions in the early 20th century, the single salary schedule was developed in order to ensure that teachers with the same years of experience and education levels received an equitable salary. By the 1950's 97% of the schools had adopted a single salary schedule (Podgursky & Springer, 2007). The same salary schedule model is still largely in effect today. In July 2005, the state of Minnesota approved Q-Comp, a performance based merit pay program for teachers. Principals do not have a designated role within the program. In some districts evaluations are conducted by teachers designated as evaluators while in other districts principals also play a role in determining compensation.

The basic premise for using teacher evaluations to determine bonuses is to help promote student achievement. Research has shown that the development of an acceptable teaching evaluation process has numerous inherent difficulties associated with it. Studies have questioned the accuracy of the rating, the subjective bias against some categories of

teachers (Sanders & Horn, 1998), the measurement of value-added assessments (Jacob & Lefgren, 2008), and inter-rater reliability (Ellett et al., 1996). Some theorists argue that teaching is not conducive to performance measures as are other professions due to the complexity of variables and the nature of the task (McCaffrey, Lockwood, Koretz, Louise, & Hamilton, 2003). This argument is countered by recent studies using longitudinal data that support value-added measurements (Jacob, 2008). Even with these advances, teaching is multidimensional and the use of student achievement data as the sole determiner of performance pay can be questioned (Peterson, 2000). Indeed, studies have shown that high stakes testing can result in undesirable consequence such as teacher cheating and test driven instruction (Jacob & Levitt, 2003). Thus, theorists support multiple measures for assessing teacher performance when the results will be used for high stake purposes such as compensation (Peterson, 2000).

Pay for performance systems have significant advantages from a cost-benefit position. With the current single salary schedule, it is nearly impossible to align performance and pay. Performance pay schemes attempt to resolve this issue. However, while there is widespread agreement that teachers do have an effect on the achievement level of students (Goe, 2007), there is not agreement on the ability for any system to determine the effect of any single teacher on the achievement levels of a group of associated students. Thus, once teachers are tenured, schools are compensating some teachers in excess of their value while under compensating others.

Researchers also speculate that the lack of incentive and constraints on wages could be a factor in the turnover of high ability teachers (Podgursky, Monroe, & Watson, 2004). This study was supported by Hoxby and Leigh (2004) who found high ability

women starting in the 1960's and continuing to the present left teaching due to the push of salary constraints rather than the pull of other employment opportunities. Theorists speculate that by aligning the compensation systems of the education profession more closely to that of other professions, high ability personnel will migrate to teaching as a profession that rewards excellence.

Pay for Performance Implications for Principals

Given the controversial issues regarding the association between value-added achievement measures and teacher effectiveness, several researchers have turned to subjective principal evaluations as an alternative for the awarding of teacher compensation (Jacob & Lefgren, 2008; Podgursky & Springer, 2007). The results of these studies suggest policy makers should consider including principal evaluations as a component of a performance pay system. Principal generated evaluation ratings were found to be a substantially better predictor of future student achievement than either experience or education, the two major components of the single salary system. Additionally, principal conducted evaluations allow the flexibility to consider the various facets of teaching in determining a final judgment rather than the narrow focus of achievement testing. A principal can consider the value teachers bring to the organization rather than merely the merit of their output (Scriven, 1987).

Incorporation of principal generated evaluation ratings into a performance pay system is not without cautionary notes. An extensive study in the 1980's (Wise et al., 1984) concluded that principals felt evaluation activities were a waste of time and a "necessary evil." While the reasoning behind the perception was not explored, high stakes evaluations for a compensation system that judiciously awards merit must rise to a

higher standard of accuracy and defensibility (Loup et al., 1996). Principals need to devote time and energy to evaluation as a serious duty that is thoughtfully executed. Indeed, Odden (2004) speculated that evaluators do become more serious when the results are used for consequential decisions. An additional concern falls within the sociological arena. Due to role conflicts and leadership power structures, principals will work to avoid conflicts and discord with their staff. Thus, elevated requirements for salary enhancements can result in inflated evaluator ratings (Loup et al., 1997). Factors such as bias and inter-rater reliability need to be addressed through training. If these factors are not addressed, previous studies have indicated the validity of the evaluations will be questioned (Heneman & Milanowski, 2003).

Summary

The review of the literature points out the following:

- Teacher evaluation has evolved, reflecting changing political and economic contexts.
- National standards for evaluating educational personnel have served as nationally recognized points of reference for evaluation system design and research.
- Approaches to evaluation vary depending upon the purpose of evaluation-formative or summative.
- The role of the principal in the evaluative process is complicated by technical and sociological aspects both of which influence a principal's ability and resolve to evaluate teachers effectively.
- Contemporary performance pay systems are influenced by labor market and technological advances in the area of value-added measurement of student achievement.

This study seeks to explore the characteristics of principals as effective evaluators for purposes of judging the quality of teachers. The literature review reveals substantial evidence principals can differentiate between highly effective and minimally effective

teachers. Furthermore, the means and impetus for quality evaluations are alluded to in the personnel evaluation standards and the performance pay movement. Yet, the literature also clearly defines teachers' perceptions as a major factor influencing the acceptability of an evaluator's rating in performance pay systems. By focusing on principals in their role as an evaluator, the study addresses a gap in the literature concerning those characteristics that contribute to the profile of an effective principal evaluator.

CHAPTER III

METHODOLOGY

Introduction

This chapter provides a description of the research design, defines the research population including the sample selection, explains the development and rationale for the research instruments, and describes the method of data collection.

The primary purpose of the study was to study the effectiveness of principals as evaluators of teachers for the explicit purpose of determining teachers' performance levels. The study defines effective evaluators as those evaluators whose practices most closely follow the personnel evaluation standards developed by the Joint Committee on Standards for Educational Evaluation (Joint Committee, 2009). The effectiveness of principals is based upon the perceptions of both teachers and principals. Because validity and reliability are of heightened importance when evaluation results are applied to performance pay decisions (Loup & Ellett, 1997), the study was placed within the context of the Minnesota Quality Compensation for Teacher Program (Q-Comp), a performance based merit pay system.

The study seeks to use perceptions of teachers and principals in Minnesota Q-Comp middle schools to answer the following questions:

1. Which practices of effective evaluators do principals follow as perceived by teachers and principals?
2. Which practices of effective evaluation are most important to teachers?
3. What is the relationship between the effective practices that principals follow and those that are most important to the teachers?

Related goals of the study include implications of the findings for district, organizational, and principal preparation programs; and the viability of using principals as primary decision makers in awarding performance pay.

Research Design

The research design involved a survey methodology. The purpose of the survey design was to make descriptive statements about the effectiveness of principals as evaluators of teachers through generalizing from a sample to a population so that inferences could be made (Cresswell, 1994). A survey was chosen for the research design based on its three advantages: the economy of the design, the rapid turnaround in data collection, and the ability to generalize characteristics of a population from a representative sampling (Creswell, 1994). The intent of the study was to obtain accurate information regarding the effectiveness of principals as evaluators in Minnesota middle schools, not to conduct a test of hypotheses.

Research Population

The population for this study included all tenured teachers who taught in Minnesota middle schools that participated in the Minnesota Q-Comp Program in the 2007-2008 school year and whose personnel evaluations conducted annually by their principal affected the awarding of their performance pay. The research population was limited to middle schools containing two or three middle grades (5 – 9) and whose principal's sole assignment was to administrate the school. The pool of subjects (Appendix A) was identified through reviewing the Q-Comp program applications for districts participating in the program in 2007-2008. These public documents were available from the Minnesota Department of Education. If an application was not clear as

to the principal's role in the awarding of performance pay, an email was sent to a representative of the district or the school principal for clarification. Finally, the principal of each potentially qualifying school was contacted to insure that the school met all the criteria.

Table 1

Population Sample of Minnesota Q-Comp Middle Schools

Total Number of Q-Comp Middle Schools (2007-2008)	43
Total number of schools with principals dedicated to only one school	41
Total number of middle schools in which the principal's evaluation affected the awarding of merit pay for tenured teachers	23
Number of middle school in which the principal annually evaluated tenured teachers	6
Qualified school(s) that did not respond	1
Qualified school that was not included due to connection with investigator	1
Total number of schools participating	4

Sample Selection

Table 1 explains how the total number of four study schools was derived. A total number of 43 middle schools were identified as participating in the Q-Comp program in the 2007-2008 school year. Of those schools 41 had principals that served only the middle school. The administrators of the other two schools had shared responsibilities for multiple schools. In 23 of the 43 schools the principal's evaluation affected the awarding of performance pay for tenured teachers. However, in 17 of the schools the tenured staff

was not evaluated annually by the principal but rather on a cycle which varied from bi-annually to once every four years. In these schools the principal's evaluation only affected the awarding of performance pay in the year in which the tenured staff was evaluated. Tenured teachers were evaluated annually by their principals in six middle schools. All of these schools had dedicated principals whose evaluations affected the awarding of performance pay on an annual basis. One of the schools was administered by the investigator of this study and, thus, was excluded from the sample due to potential conflict of interest. The superintendents of all five remaining districts responded positively to inclusion in the study. However, one middle school principal elected not to have her school participate through implying a negative response by not responding to inquiry emails and phone messages. The net result was the participation of four schools out of a total population of five schools (80%).

Table 2

Population Sample of Teachers and Principals

Population	Tenured Staff	Principals
Middle School #1	31	1
Middle School #2	32	1
Middle School #3	22	1
Middle School #4	46	1
Totals	131	4
Total number survey respondents	76	4

Table 2 explains how the total number of study participants was derived. The total population of 131 was obtained by totaling the number of tenured staff from each building. Building principals provided the totals of tenured staff in each of their respective buildings. Because the study design assured staff that individual and building results would be completely anonymous, no building identifiers were placed on responses. Thus, the total number of respondents per building is not known. The on-line survey link was forwarded to tenured staff through the building principals insuring all tenured staff received the emails with the survey link. Reminder emails were also sent. A total of 55 tenured teachers elected not to participate in the study through implying a negative response by not taking the survey. The net result was a total of 76 teachers out of 131 in the total population (58%) based on the four participating schools. Table 2 also explains the total number of principal participants. Each school had one principal. While the individual responses were anonymous, a total of four respondents completed the principal survey. This on-line survey link was only emailed to the principals of the participating schools. The net result was a total of four principals out of four in the total population (100%). See Table 3 for a summary of the response rate. Appendix A provides a list of all the schools that met the criteria for the study and a summary of those schools that participated in the study.

Table 3

Summary of the Response Rate of Study Participants

Emails to Study Participants	Number of Responses	Percentage Rate of Responses
Tenured teachers	76	58% (76-131)
Principals	4	100% (4-4)

Research Instruments

Two survey research instruments, one for the principal survey and one for the teacher survey, were employed for this study (Appendix B). The surveys were piloted in the fall of 2008 and distributed on-line to study participants in the winter of 2009. The on-line surveys provided an efficient, timely format for distribution. When supported with appropriate reminders, a web-based survey has consistent findings with traditional methods (Gosling, Vazire, Strivastava, & John, 2004) and should have adequate response (Kaplowitz, Hadlock & Levine, 2004).

Both the principal survey and the teacher survey consisted of 24 identical statements of effective practice drawn from the personnel evaluation standards. These 24 statements served as a 24-item scale that defined effective evaluation practices for the study. The 24-item scale used a 5-point Likert scale anchored by “never” and “always” to determine the frequency by which the practice was followed. Those practices with the highest reported frequencies by both teachers and principals were determined to be the effective practices followed by principals. The same 24-item scale was also used in the teacher survey to measure the relative importance of each practice. The response for this application was a 3-point Likert scale anchored by “not important” and “important.”

Principal respondents were asked to read each statement and scale their perceptions of the degree to which their evaluation practices corresponded to the statement. The second part of the principal survey asked respondents to provide

demographic data regarding: (1) gender, (2) age, and (3) number of years of principal experience.

The teacher survey instrument asked the participants to respond to the 24-item scale based upon the practices of the most competent supervising principal in the area of teacher evaluation they have had in their teaching career. Respondents were asked to read each statement and scale their perceptions of the degree to which their most competent supervising principal's evaluation practices were reflected in the statement. The degree of similarity was quantified using a 5-point, Likert scale. The second part of the survey asked the teacher respondents to indicate the importance of each item as a factor in an effective evaluation. The degree of importance was quantified using a 3-point, Likert scale. The third part of the teacher survey asked respondents to provide demographic data regarding: (1) gender, (2) age, and (3) educational background.

Development of the Research Instruments

The instruments were developed utilizing the *Personnel Evaluation Standards* (Joint Committee, 2009). *The Personnel Evaluation Standards* groups the standards into four attributes: *propriety*, *utility*, *feasibility*, and *accuracy*. These four attributes are considered essential qualities of sound evaluation practice. In addition, an explanation and a list of guidelines that give evaluators suggested procedures to help meet the requirements of the standard is included for each standard. *The Personnel Evaluation Standards* also contains a Functional Table of Contents which further organizes the standards by common uses such as training and certification. These elements were essential in the development of the instrument and the organization of the data.

The research instrument was developed in two stages. The first stage involved constructing the 24-item scale using the personnel standards and verifying that the practices included in the survey were under the decision making authority of principals and not district policy.

To create the 24-item scale, seven standards were selected as most pertinent to the study. The selection of the seven standards was based upon (1) the standards' designation in the Functional Table of Contents of *The Personnel Evaluation Standards* (2009) as most applicable to merit pay, tenure, or promotion decision making; and (2) through the identification of the standards by a focus group of principals as being under a principal's decision making authority as opposed to district policy. Next, 24 statements of effective practice were drawn from the guidelines of the seven standards. These 24 statements served as the 24-item scale that defined effective evaluation practices for the study. The number of scale items was not consistent across standards as some standards' guidelines addressed principals' practices more than others. Similarly, the number of standards was not consistent across the four attributes. The principal focus group also reviewed the 24 statements insuring they were associated with principal practices and not district policy.

Table 4

Relationship of Survey Scale Statements to Personnel Evaluation Standards

Attribute	Standard Code	Standard	No. of Statements on Instrument
Propriety	P4	Interactions with Employees	6
	P5	Comprehensive Evaluation	5
Utility	U4	Explicit Criteria	3
Feasibility	F1	Practical Procedures	2

Accuracy	A1	Valid Judgments	3
	A8	Bias Identification and Management	3
	A10	Justified Conclusions	2

The second stage of the survey involved piloting the surveys electronically with both principals and teachers. The pilots were conducted with schools in the investigator's district. Feedback on the instruments was gathered both electronically and through focus groups. Questions were screened for clarity.

Method of Data Collection

The survey of the middle school teachers and principals was conducted in the winter of 2009. The total population of principals and teachers was identified through Minnesota Department of Education documentation and the investigator's personal contact with districts prior to data collection. Listed below are the steps describing the method of data collection:

1. Superintendent letter- Superintendents of districts with subject schools were sent a letter on January 25, 2009 stating the purpose and intent of the study and requesting their consent to allow the investigator to contact the principal(s) of the subject school(s) in their district (see Appendix C). Superintendents were asked to return a form indicating their approval to contact the school(s).

2. Principal letter- Principals of the subject schools were sent a letter electronically requesting their participation in the study. The purpose and intent of the study was stated. The email also included an electronic copy of the principal and teacher survey. It was explained that communication would be via email and the survey would be conducted on-line via web-based survey software. After one week, follow up emails and phone calls were made to principals that had not responded.

3. Principal surveys- Principals were sent an email letter with specific instructions regarding how to complete the on-line survey. A link to the survey was also included in the email letter. Principals were given three weeks to complete the survey.

4. Teacher surveys- Email letters stating the purpose, intent and confidentiality of the study was sent to the principals to be forwarded to their tenured staff. The email letter also contained a link to the on-line survey with instructions for completing the survey. Teachers were given three weeks to complete the survey. Reminder emails for teachers were sent to the principals to be forwarded to their staffs.

5. Final communication- A final email was sent to principals to be forwarded to their staffs. This email thanked staff and principals for their participation.

Given the purpose of the research was to study the effectiveness of principals as evaluators of teachers based on whether or not the principals' evaluation practices reflect the personnel evaluation standards, the survey instrument sufficed as a means to gather pertinent data. As perceived by principals and teachers, principals either followed the practices of the standards or they did not. More descriptive data acquired through focus groups and interviews could be included in future studies.

CHAPTER IV

RESULTS

Introduction

This chapter provides an analysis of the data gathered from the principal and teacher evaluation surveys distributed to middle school principals and tenured teachers who participated in the Minnesota Q-Comp program in 2007-2008. A *t*-test statistical analysis was applied to the mean values of the responses to determine the perceived frequency with which principals exemplified effective evaluation practices and the relative importance of the practices as perceived by teachers. The first section of this chapter presents demographic data regarding the study participants. The second section contains an analysis of the survey results as they apply to the research questions.

Demographic Data Analysis

Teacher Survey Participants

Seventy-six teachers participated in the survey representing 58% of the sample population. Of the teacher survey participants, 21 were male (30%) and 49 were female (70%). The majority of respondents were between 35 and 39 years of age (19.1%) with the second largest age group being teachers between 30 and 34 years of age (17.6%). The smallest age group of participants was teachers 29 years of age or younger (7.4%). The large majority of teachers held master's degrees (78%) followed by those with a bachelor's degree (21%) and those with a specialist or 6th year certificate (2.8%). Those teachers who selected "other" for educational background (5.6%) indicated a variety of specialist certificates.

Table 5

Characteristics of Teacher Participants

Gender			Age			Educational Background		
	n	%		n	%		n	%
Male	21	30	29 or below	5	7.4	B.A. or B.S	15	21.1
Female	49	70	30–34	12	17.6	M.A. or Ed.M.	56	78.9
-	-	-	35–39	13	19.1	Ed Sp	2	2.8
-	-	-	40–44	6	8.8	-	-	-
-	-	-	45–49	8	11.8	-	-	-
-	-	-	50–54	9	13.2	-	-	-
-	-	-	55–59	6	8.8	-	-	-
-	-	-	60+	9	13.2	-	-	-

Principal Survey Participants

Four principals participated in the survey representing 100% of the sample population. The principal participants were gender balanced with two males and two females. The majority of the principals were 55 years of age or older with one principal between 55–59 years of age, one principal between 44–49 years of age, and two principals being 60 years old or older. The majority of principals (3) had been principals for more than 15 years. The other principal had been a principal for 6–10 years.

Table 6

Characteristics of Principal Participants

Gender			Age			Years as Principal		
	n	%		n	%		n	%
Male	2	50	29 or below	0	0	1–5	0	0
Female	2	50	30–34	0	0	6–10	1	75
-	-	-	35–39	0	0	11–15	0	0
-	-	-	40–44	1	25	15+	3	75
-	-	-	45–49	0	0	-	-	-
-	-	-	50–54	0	0	-	-	-
-	-	-	55–59	1	25	-	-	-
-	-	-	60+	2	50	-	-	-

Statistical Analysis by Effective Practices

The study sought to answer three questions:

1. Which practices of effective evaluators do principals follow as perceived by teachers and principals?
2. Which practices of effective evaluation are the most important to teachers?
3. What is the relationship between the effective practices that principals follow and those that are most important to the teachers?

To answer the questions, practices of effective evaluators was defined as those practices which follow the personnel evaluation standards as outlined in *The Personnel Evaluation Standards* (Joint Committee, 2009). The personnel standards are grouped into four attributes: *propriety*, *utility*, *feasibility*, and *accuracy*. The Joint Committee also

provided a list of guidelines for each standard that gives evaluators suggested procedures to help meet the requirements of the standards. These guidelines were used to develop the 24-item scale for both survey instruments. Of the 27 personnel evaluation standards, seven standards were selected as most pertinent to the study. Throughout the data analysis standards are identified by standard codes where the first letter indicates the attribute (P = *propriety*, U = *utility*, F = *feasibility*, A = *accuracy*) followed by the standard number. Thus the standard code P4 represents the fourth *propriety* standard. The results of the study are analyzed by standard using averages and ranking. After an analysis by standard, the results are summarized by attribute. Analysis will address frequency data, importance data, and the relationship between the data.

Analysis by Standard

Frequency of Application of Effective Practices by Principals

The scale statements in the survey instruments were drawn from practices of effective evaluators. The scores for each statement were averaged and ranked on both the teacher and principal surveys. Practices ranked high in frequency on both surveys were considered frequently followed by principals. Table 7 lists the average score and ranking by both teachers and principals. Eight statements fell in the top nine rankings of both the teachers and the principals. The statements are listed in Table 7 with their associated personnel standard code.

Table 7

Average Score and Ranking of Each Statement

Statement No.	Standard Code	Statement Text	Teacher Average	Principal Average	Teacher Rank	Principal Rank
1	**P4**	**Conducted the evaluation in a respectful manner.**	**4.626667**	**5**	**2**	**1**
2	**P4**	**Conducted the evaluation in a manner that I considered fair.**	**4.486486**	**5**	**3**	**1**
3	**P4**	**Communicated the results clearly and objectively.**	**4.391892**	**4.75**	**4**	**5**
4	P4	Developed a relationship of mutual trust and understanding with me prior to the evaluation.	4.121622	4.5	10	12
5	**P4**	**Demonstrated a genuine interest in me as a person.**	**4.148649**	**5**	**9**	**1**
6	P4	Took into account my personal and professional needs.	4.013514	4.75	16	5
7	P5	Ensured I knew what would be assessed.	4.222222	4.25	7	16
8	P5	Ensured I knew how the evaluation data would be collected.	3.985714	4.5	17	12
9	P5	Described and justified the basis for interpretation of both positive and negative assessment information and results.	4.098592	4.5	12	12
10	**P5**	**Reported fully both strengths and weaknesses with supporting evidence.**	**4.236111**	**4.75**	**6**	**5**
11	P5	Supported continued, timely professional growth.	4.069444	4.75	13	5
12	U4	Addressed only identified professional roles and responsibilities in the evaluation report and ensured that extraneous comments beyond the criteria were neither included nor accepted.	4.112676	3.75	11	22
13	**U4**	**Provided copies of written evaluation reports to me.**	**4.638889**	**5**	**1**	**1**
14	**U4**	**Used the agreed upon criteria, providing a rationale and justification of evaluation findings.**	**4.388889**	**4.75**	**5**	**5**
15	F1	Delineated the procedures by which I could exercise my rights to review data about my performance.	3.380282	4.25	21	16
16	F1	Encouraged me to suggest ways by which evaluation procedures could be made more efficient and useful.	2.742857	2.75	24	24
17	A1	Considered my background and cultural experiences when interpreting performance.	3.521739	4.25	20	16
18	A1	Ensured that my scoring was not influenced by factors irrelevant to my performance being evaluated (general impression or previous rating influences the present rating).	3.873239	4.25	18	16
19	A1	Ensured that my summary conclusions which were derived from a series of assessments corresponded with the documented results.	4.059701	4.75	14	5
20	A8	Obtained data and judgments from multiple sources to ensure validity and consistent indications of my performance.	3.352113	3.75	22	22

21	A8	Allowed me to review data and participate in interpreting it where appropriate.	4.044118	4	15	21
22	**A8**	**Based my evaluations on defensible information with conclusions that were justifiable.**	**4.214286**	**4.75**	**8**	**5**
23	A10	Generated, assessed, and reported plausible alternative explanations of findings and, if appropriate, indicated why these explanations should be discounted.	3.15625	4.25	23	16
24	A10	Limited conclusions to those situations, time periods, contexts, and purposes for which the evaluation findings were applicable.	3.867647	4.5	19	12

Note. Bold items are higher ranked in frequency by both teachers and principals

A frequency analysis was conducted to identify statements in which principals rated the frequency of their practice higher than the teachers scored principals. A one-sided *t*-test with a confidence level of 95% ($p < .05$) was performed to test this hypothesis. Nine statements showed a significant difference between the principals' and the teachers' scores.

Table 8

Frequency of Effective Practices between Teachers and Principals

					Frequency	
					Teachers	Principals
Statement Number	Standard Code	p-value	t-score	df	Average Answer	Average Answer
1	**P4**	**0.000**	**-5.117**	**74**	**4.627**	**5**
2	**P4**	**0.000**	**-6.248**	**73**	**4.486**	**5**
3	P4	0.125	-1.362	3.665	4.392	4.75
4	P4	0.144	-1.234	3.819	4.122	4.5
5	**P4**	**0.000**	**-7.861**	**73**	**4.149**	**5**
6	**P4**	**0.026**	**-2.698**	**4.26**	**4.014**	**4.75**
7	P5	0.462	-0.102	4.236	4.222	4.25
8	P5	0.087	-1.653	4.404	3.986	4.5
9	P5	0.132	-1.302	3.901	4.099	4.5
10	P5	0.064	-1.914	3.986	4.236	4.75
11	**P5**	**0.032**	**-2.496**	**4.241**	**4.069**	**4.75**
12	U4	0.636	0.381	3.086	4.113	3.75
13	**U4**	**0.000**	**-4.389**	**71**	**4.639**	**5**
14	U4	0.125	-1.348	3.946	4.389	4.75
15	**F1**	**0.013**	**-2.952**	**5.748**	**3.380**	**4.25**
16	F1	0.495	-0.014	3.692	2.743	2.75
17	A1	0.114	-1.45	3.639	3.522	4.25
18	A1	0.120	-1.344	4.727	3.873	4.25
19	**A1**	**0.030**	**-2.489**	**4.553**	**4.060**	**4.75**
20	A8	0.111	-1.384	5.224	3.352	3.75
21	A8	0.528	0.075	3.319	4.044	4

22	A8	0.059	-1.969	4.205	4.214	4.75
23	**A10**	**0.005**	**-3.71**	**5.76**	**3.156**	**4.25**
24	A10	0.056	-2.009	4.235	3.868	4.5

Note. Bold items are statistically different.

Relative Importance of Specific Effective Evaluation Practices

The second research question addresses the relative importance of each of the effective practices. Using the teacher survey instrument, the teachers indicated the degree of importance of each statement by responding to a 3-point Likert scale anchored by "not important" and "important." The responses to each statement were averaged and ranked. Table 9 lists the average score and ranking of each statement.

Table 9

Average Score and Ranking of Statements by Teachers

Statement Number	Standard Code	Statement Text	Average Answer	Importance Rank
1	**P4**	**Conducted the evaluation in a respectful manner.**	**2.888888889**	**2**
2	**P4**	**Conducted the evaluation in a manner that I considered fair.**	**2.907407407**	**1**
3	**P4**	**Communicated the results clearly and objectively.**	**2.888888889**	**3**
4	**P4**	**Developed a relationship of mutual trust and understanding with me prior to the evaluation.**	**2.833333333**	**5**
5	P4	Demonstrated a genuine interest in me as a person.	2.518518519	18
6	P4	Took into account my personal and professional needs.	2.759259259	10
7	P5	Ensured I knew what would be assessed.	2.740740741	11
8	P5	Ensured I knew how the evaluation data would be collected.	2.555555556	16
9	**P5**	**Described and justified the basis for interpretation of both positive and negative assessment information and results.**	**2.833333333**	**6**
10	**P5**	**Reported fully both strengths and weaknesses with supporting evidence.**	**2.87037037**	**4**
11	P5	Supported continued, timely professional growth.	2.666666667	14
12	U4	Addressed only identified professional roles and responsibilities in the evaluation report and ensured that extraneous comments beyond the criteria were neither included nor accepted.	2.648148148	15

13	U4	Provided copies of written evaluation reports to me.	2.796296296	9
14	**U4**	**Used the agreed upon criteria, providing a rationale and justification of evaluation findings.**	**2.814814815**	**7**
15	F1	Delineated the procedures by which I could exercise my rights to review data about my performance.	2.518518519	19
16	F1	Encouraged me to suggest ways by which evaluation procedures could be made more efficient and useful.	2.166666667	23
17	A1	Considered my background and cultural experiences when interpreting performance.	2.018518519	24
18	A1	Ensured that my scoring was not influenced by factors irrelevant to my performance being evaluated (general impression or previous rating influences the present rating).	2.555555556	17
19	A1	Ensured that my summary conclusions which were derived from a series of assessments corresponded with the documented results.	2.703703704	12
20	A8	Obtained data and judgments from multiple sources to ensure validity and consistent indications of my performance.	2.444444444	21
21	A8	Allowed me to review data and participate in interpreting it where appropriate.	2.685185185	13
22	**A8**	**Based my evaluations on defensible information with conclusions that were justifiable.**	**2.814814815**	**8**
23	A10	Generated, assessed, and reported plausible alternative explanations of findings and, if appropriate, indicated why these explanations should be discounted.	2.277777778	22
24	A10	Limited conclusions to those situations, time periods, contexts, and purposes for which the evaluation findings were applicable.	2.5	20

Note. Bold items are statistically important

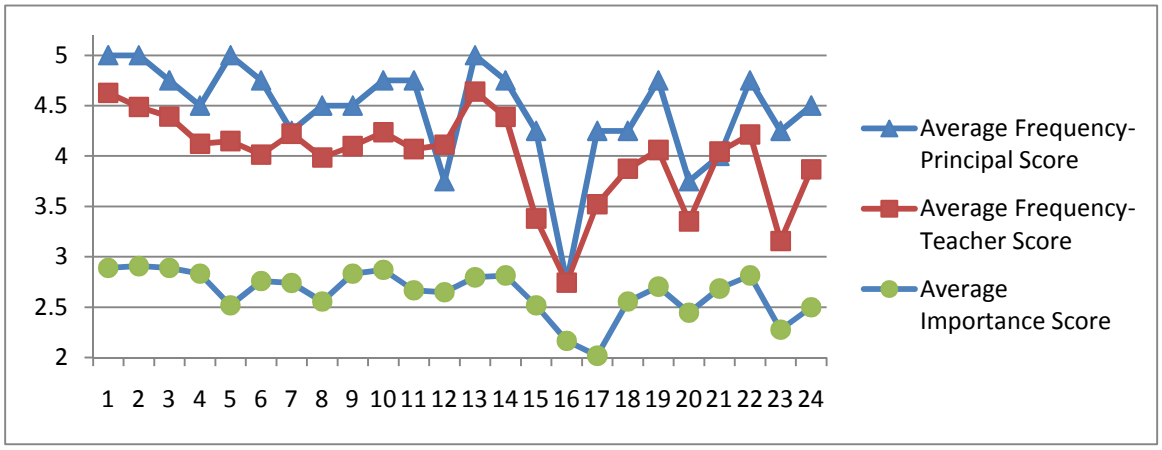
A two-sided *t*-test was applied to determine which statements were significantly most important. Appendix D shows the results of this analysis. For each statement, the total number of statements statistically less important than that statement was determined. Eight statements were identified (#1, #2, #3, #4, #9, #10, #14 and #22) which are the most significant. Of these, statements two and three were found to be the most important as they were more significantly important than 15 other statements.

Frequency and Importance of Effective Practices

An examination of the average frequency scores by teachers and principals reveals that, with the exception of statement 12, principals consistently score themselves higher in frequency of applying effective practices than the teachers score the principals. Although the levels of importance for each statement was rated on a 3-point Likert scale instead of the 5-point Likert scale used for frequency, data show a general parallel trend with higher frequencies reported on the statements that are also rated higher in importance. Conversely, lower frequencies are reported by both teachers and principals on statements rated lower in importance.

Figure 1

Average Frequency and Importance Scores



Statistical Analysis by Attribute

The statements for the survey were derived from seven personnel standards. Each of the four attributes identified by the Joint Committee (2009) as essential components of effective evaluation was represented by at least one of the seven standards. Table 10 provides A summarizing of data by attribute and standard provides a deeper

understanding of principals' practices in teacher evaluation as perceived by both teachers and principals.

Propriety Standards

The focus of *propriety standards* is to insure that evaluations are conducted legally, ethically, and with proper concern for the welfare of those involved in the evaluation (Joint Committee, 2009). Two *propriety standards* were incorporated into the study- Standard P4 (Interactions with Evaluatees) and Standard P5 (Comprehensive Evaluation). Six scale statements were developed from Standard P4 and five scale statements were developed from Standard P5.

Standard P4 (Interactions with Evaluatees) scored high in both frequency and importance. Four of the six scale statements from Standard P4 were ranked in the top five for frequency by the teachers while three of the six were ranked in the top five by the principals. Four of the six statements (#1, #2, #5, and #6) also showed a significant difference in the frequency score between teachers and principals with principals ranking their frequency significantly higher. Teachers ranked four of the six Standard P4 statements in the top five for importance. The same four statements are also found to be statistically more important statements as determined by a two-sided *t*-test.

Standard P5 (Comprehensive Evaluation) had mixed results in terms of frequency and importance. Only one of the five Standard P5 statements (#10) ranked relatively high in frequency with the teachers and the principals with a 6^{th} and 5^{th} ranking respectively out of 24. The other four Standard P5 statements (#7, #8, #9, and #11) ranked in the middle third for frequency. Principals scored themselves significantly higher on one (#11) of the five statements. Two of the five statements (# 9 and #10) ranked in the top

third for importance and were statistically important. Two statements (#7 and #11) ranked in the middle third while one statement (#8) ranked in the lower third for importance.

Utility Standards

The focus of *utility standards* is to insure that evaluations are conducted in a manner that insures evaluations are timely, informative, and useful (Joint Committee, 2009). One utility standard was incorporated into the study- Standard U4 (Explicit Criteria). Three scale statements were developed from Standard U4.

Standard U4 (Explicit Criteria) also had mixed results in terms of frequency and importance. Two of the three Standard U4 statements (#13 and #14) ranked high in frequency with the teachers and the principals ranking the statements 1st and 5th respectively. Principals rated themselves significantly higher in frequency on Standard U4 statement #13. The two statements also ranked in the top third for importance with #14 being a statistically significant important statement. The other Standard U4 statement (#12) ranked in the middle third for frequency and in the middle third for importance.

Feasibility Standards

The focus of *feasibility standards* is to insure that evaluations are easily implemented, efficient in terms of time and resources, and are adequately funded (Joint Committee, 2009). One feasibility standard was incorporated into the study- Standard F1 (Practical Procedures). Two scale statements were developed from Standard F1.

Standard F1 ranked lower in terms of frequency and importance. Both of the Standard F1 statements (#15 and #16) were ranked in the lower third in frequency by both teachers and the principals. Statement #15 showed a significant difference in the

frequency score between teachers and principals with principals ranking their frequency significantly higher. The two statements also ranked in the lower third for importance.

Accuracy Standards

The focus of *accuracy standards* is to insure that evaluations are technically adequate so that the information generated can be used to make sound judgments (Joint Committee, 2009). Three accuracy standards were incorporated into the study- Standard A1 (Valid Judgments), Standard A8 (Bias Identification and Management), and Standard A10 (Justified Conclusions). Three scale statements were developed from Standard A1, three scale statements were developed from Standard A8, and two scale statements were developed for Standard A10.

Standard A1 (Valid Judgments) ranked lower in terms of frequency and importance. Two of the Standard A1 statements (#17 and #18) were ranked in the lower third in frequency by both teachers and the principals. Statement #19 ranked in the middle third for both frequency and importance. Statement #19 also showed a significant difference in the frequency score between teachers and principals with principals ranking their frequency significantly higher.

Standard A8 (Bias Identification and Management) had mixed results in terms of frequency and importance. Two of the Standard A8 statements (#20 and #21) were ranked in the lower third in frequency by both teachers and principals. Statement #20 also ranked in the lower third for importance. Statement #21 ranked in the middle third for importance. Statement #22 ranked in the top third for both frequency and importance.

Standard A10 (Justified Conclusions) was lower ranked in terms of frequency and importance. One of the Standard A10 statements (#23) was ranked in the lower third in

frequency by both teachers and principals. Statement #23 showed a significant difference in the frequency score between teachers and principals with principals ranking their frequency significantly higher. Statement #23 ranked in the lower third for importance. Statement #24 ranked in the middle third for frequency by both teachers and principals. Statement #24 ranked in the bottom third for importance.

Analysis of data by standards indicates that standards which are of greater relative importance are generally perceived by both teachers and principals as being followed more frequently. Generally speaking *propriety standards* are deemed the most important and *accuracy standards* the least important. *Propriety standards*, especially Standard P4 (Interactions with Evaluatees), is highly valued and followed. The *feasibility standard* and *accuracy standards*, most notably Standard A8 (Bias Identification and Management) are less valued and followed.

CHAPTER V

DISCUSSION

Introduction

This chapter discusses the findings and implications of the study. The chapter is organized into the following sections: (a) summary of the purpose of the study, (b) significance and rationale for the study, (c) review of the procedures, (d) review of the main findings, (e) discussion of the conclusions, (f) recommendations, and (g) suggestions for further research.

Summary of Purpose

The primary purpose of this study was to gather perceptions of principals and teachers with regard to the effectiveness of principals as evaluators of teachers. Perceptions were reviewed within the context of seven standards across the four attributes of the personnel standards developed by the Joint Committee on Standards for Educational Evaluation (Joint Committee, 2009). These standards, organized by attributes, define quality personnel evaluation in education. Principal and teacher participants were asked to identify the frequency with which principals followed effective evaluation practices. Teacher participants were also asked to identify the relative importance of the practices as a factor of effective evaluation. The study used perceptions of middle school principals and tenured teachers who participated in the Minnesota Q-Comp program in 2007-2008 to answer the following research questions:

1. Which practices of effective evaluators do principals follow as perceived by teachers and principals?
2. Which practices of effective evaluation are most important to teachers?
3. What is the relationship between the effective practices that principals follow and those that are most important to the teachers?

Significance and Rationale of the Study

Several features of the study are unique. The data from the study provide feedback in regard to the perceived effectiveness of principals as evaluators of teachers based on nationally established standards for personnel evaluation. Other studies have attempted to determine the accuracy of principal evaluation based on student and parent surveys (Peterson, 2005, 2000; Wilkerson et al., 2000) or achievement data (Jacob, 2008; Medley & Coker, 1987) but no study could be located that based evaluation effectiveness on the frequency by which standards of evaluative practice are followed. Furthermore, the study also provides data on the relative importance of each practice as perceived by the teachers, thus allowing for an effectiveness analysis based on not only the frequency by which a practice is followed but also on the relative importance of that practice as an element of effective evaluation.

The study also focuses on tenured teacher evaluation. The literature highlights the more skeptical nature of tenured teachers with regard to personnel evaluations (Kauchak et al., 1985). While the need for effective evaluation of tenured teachers (Bridges, 1992; Tucker, 1997) is cited in literature, few studies address only tenured teacher evaluation.

Finally, there is little research on the effectiveness of principals' evaluation practices using data gathered within the context of a performance pay system such as Q-Comp. In these programs, literature suggests the competency of the evaluator and the resulting validity of the evaluation becomes more critical (Loup & Ellett, 1997; Loup et al., 1996). The Minnesota Q-Comp program provides a high stakes context for gathering data on the teachers' and principals' perceptions.

Review of Procedures

The study employed a quantitative methodology which assessed the degree of effectiveness of principals' evaluations through a teacher survey and a principal survey. Both instruments featured a list of 24 identical statements of effective evaluation practices derived from personnel standards (Joint Committee, 2009). The surveys, distributed on-line, involved principals and tenured teachers in four middle schools that participated in the Minnesota Q-Comp program in 2007-2008. In both surveys principals and teachers indicated the degree to which they perceived principals followed the effective evaluation practices. In addition, teachers indicated their perception of the relative importance of each practice as it related to effective teacher evaluation. A total of 76 teachers and 4 principals responded to the surveys. The surveys were developed in the fall of 2008 and were conducted in the winter of 2009.

Review of Main Findings

Question 1: What practices of effective evaluators do principals follow as perceived by teachers and principals?

Practices Ranked High in Frequency

Teachers and principals tended to agree on those practices most frequently followed by principals. The practices that were ranked high in frequency by both teachers and principals were primarily from the *propriety attribute* which addresses ethical considerations and regard for the welfare of the evaluatee. The highest rankings in frequency were found in practices linked to Propriety Standard P4 (Interactions with Evaluatees). The six practices linked to Propriety Standard P4 were (a) conducting the evaluation in a respectful manner, (b) conducting the evaluations in a fair manner, (c)

communicating the results clearly and objectively, (d) developing a relationship of trust prior to evaluation, (e) demonstrating a genuine personal interest in the evaluatee, and (f) taking into account personal and professional needs. Of the six practices, four were identified as high frequency by both teachers and principals. These included (a) conducting the evaluation in a respectful manner; (b) conducting the evaluations in a fair manner, (c) communicating the results clearly, and (d) demonstrating a genuine personal interest in the evaluatee. While both teachers and principals identified these practices as high frequency, principals significantly ranked themselves at a higher frequency in (a) conducting evaluations in a respectful manner, (b) conducting evaluations in a fair manner, and (c) demonstrating a genuine interest in the person being evaluated.

The *utility attribute* addresses the usefulness and timeliness of evaluation. Utility Standard U4 (Explicit Criteria) had two practices both of which were ranked high frequency by teachers and principals. These included (a) providing copies of evaluations and (b) using agreed upon criteria. In one of the practices (providing copies of evaluations) principals significantly ranked themselves at a higher frequency than teachers.

The *accuracy attribute* addresses the technical accuracy and completeness of the evaluation. Accuracy Standard A8 (Bias Identification and Management) had one practice (using defensible information) which was identified as high frequency by both teachers and principals. The rating of frequency in this practice was not significantly different between teachers and principals.

Practices Ranked Low in Frequency

Teachers and principals also tended to agree on those practices least frequently followed by principals. Not as many of the rankings in the least frequent practices were significantly different between teachers and principals as opposed to the high frequency practices in which this commonly occurred. The practices associated with the *feasibility* and *accuracy attributes* tended to be ranked the lowest in frequency by both teachers and principals.

The *feasibility attribute* addresses the ensured ease of implementation and efficiency. Only Feasibility Standard F1 (Practical Procedures) was used in the study. Both of the two practices linked with this standard (teachers' rights to review data and use of agreed upon criteria with justification of findings) were ranked low in frequency by teachers and principals. While ranked low in frequency overall, principals significantly ranked themselves at a higher frequency in one practice (teachers' rights to review data).

The three standards associated with the *accuracy attribute* included in the study were (a) Accuracy Standard A1 (Valid Judgments), (b) Accuracy Standard A8 (Bias Identification and Management), and (c) Accuracy Standard A10 (Justified Conclusions). The practices linked to these standards tended to be ranked low in frequency. The three practices associated with Accuracy Standard A1 were (a) consideration of background in interpreting performance, (b) insuring scoring was not influenced by irrelevant factors, and (c) insuring summary conclusions were derived from a series of assessments. Of the three practices, two (consideration of background in interpreting performance and insuring scoring was not influenced by irrelevant factors) were low ranked in frequency

by both teachers and principals. Neither of the two practices showed a significant difference in rating between teachers and principals. Principals did rank themselves significantly higher in the third practice (insuring summary conclusions were derived from a series of assessments). Accuracy Standard A8 had one practice (obtaining data and judgments from multiple sources to insure validity) linked to it that was ranked low in frequency. Principals did not rank themselves significantly higher in this practice. The two practices linked to Accuracy Standard A10 were (a) providing plausible alternative explanations of findings and (b) limiting conclusions to applicable situations and time periods. The first practice (providing plausible alternative explanations of findings) was ranked low in frequency by both teachers and principals. However, within that low ranking, principals significantly ranked themselves at a higher frequency.

Principals and teachers generally agreed on which practices principals followed and did not follow. Principals also tended to rank themselves as more frequently following evaluation practices than did the teachers. Standards associated with the *propriety attribute* were followed more frequently than other standards.

Question 2: What practices of effective evaluation are most important to teachers?

Eight evaluation practices were significantly more important than the other practices. The eight significantly important practices included (a) conducting the evaluation in a respectful manner, (b) conducting the evaluation in a manner considered fair, (c) communicating the results clearly and objectively, (d) developing a relationship of trust prior to evaluation, (e) describing and justifying the basis for interpretation of both positive and negative assessments, (f) reporting fully both strengths and weaknesses

with supporting evidence, (g) using agreed upon criteria, and (h) using defensible information.

The majority of the most significantly important practices were linked to the *propriety attribute*. Propriety Standard P4 (Interactions with Employees) contained four of the top five important practices which were (a) conducting the evaluation in a respectful manner, (b) conducting the evaluation in a manner considered fair, (c) communicating the results clearly and objectively, and (d) developing a relationship of trust prior to the evaluation. Propriety Standard P5 (Comprehensive Evaluation) had one practice (reporting fully both strengths and weaknesses with supporting evidence) which ranked in the top five of significantly important practices.

The practices that were less important in effective evaluation tended to be associated with the *accuracy attribute*. Three of the five least important practices were from *accuracy* associated standards. These included (a) considering background and cultural experiences when interpreting performance, (b) obtaining data and judgments from multiple sources, and (c) generating plausible alternative explanations of findings. The two other least important practices were (a) encouraging suggestions of ways by which evaluation procedures could be made more efficient and useful (*feasibility attribute*) and (b) demonstrating a genuine interest in the evaluatee (*propriety attribute*).

Teachers perceived the standards associated with the *propriety attribute* as most important. These standards addressed relationships, ethical and legal issues.

Question 3: What is the relationship between the effective practices that principals follow and those that are most important to the teachers?

Those effective practices frequently followed by principals tended to also be the most significantly important practices as perceived by the teachers. Of the eight significantly important practices identified by teachers, six of the practices were also identified as high frequency practices. While both teachers and principals identified these practices as high frequency, principals significantly ranked themselves at a higher in frequency in two of the six practices: (a) conducting the evaluation in a respectful manner and (b) conducting the evaluation in a manner considered fair.

Those effective practices least frequently followed by principals tended to also be the least significantly important practices as perceived by the teachers. Two *propriety* associated practices had a variation in importance and frequency rankings. One practice (describing and justifying the basis for interpretation of both positive and negative assessments) was ranked as significantly important by teachers but was not ranked high in frequency by teachers or principals. Conversely, the other practice (demonstrating a genuine interest in the person being evaluated) was ranked high in frequency by both teachers and principals yet was not significantly important to teachers. On this practice principals ranked themselves at a significantly higher frequency. The four other lowest ranked practices in importance were also low ranked in frequency by both teachers and principals. While ranked low in frequency by teachers and principals, principals ranked themselves significantly higher in frequency on one of the four least important practices (generating plausible alternative explanations of findings).

The standards that were perceived to be most frequently followed tended to be the standards that were also identified as most important. Standards that were not as frequently followed tended to be ranked less important as perceived by the teachers.

Conclusion

If the assumption is that principals who follow evaluation practices based on the personnel standards are effective evaluators, then the data indicate that principals are effective evaluators of teachers. While principals consistently and sometimes significantly ranked themselves higher in the frequency by which they followed the practices of effective evaluation, the teachers' average frequency score on 16 of the 24 practices was above 4.0 (Often) on the 5-point Likert scale. Indeed, only one practice of the twenty-four had a teacher average frequency score lower than 3.0 (Occasionally).

The importance of the practices to teachers adds deeper insights into the conversation. Principals' practices which are ranked higher in frequency are the practices that teachers care about the most. Principals' practices that are ranked lower in frequency are of lower importance to the teachers. The only notable, significant exception is the lack of relative importance teachers placed on principals developing a genuine interest in them as a person. This practice was ranked high in frequency by both teachers and principals. Of those practices which were significantly important to teachers and reached the significant level for frequency, the lowest average frequency score was 4.486.

Not all personnel standards are of equal importance. Teachers tend to place high importance on practices that insure evaluations are conducted ethically and with due regard for the welfare of those being evaluated. They place high importance on the interactions of principals with employees. Respectfulness, perceived fairness, and clear communication are paramount to teachers in the evaluation process. Generally, they are less concerned about efficiency, practicality, or aspects that promote valid judgments to minimize misinterpretation. The one exception is a high value placed on the use of

defensible information with justifiable conclusions. Thus, if the effectiveness of the principals as evaluators is in the eyes of the beholder then the parallel rankings between importance and frequency should strengthen the argument that principals are effective evaluators of teachers as perceived by teachers and principals.

The results of the study appear to be contrary to commonly held beliefs that accuracy is critical to effective evaluation in high stakes evaluations such as those found within the context of performance pay systems (Jacob & Lefgren, 2008). One possible explanation might lie in the implementation of the Q-Comp program in Minnesota. While the program directs districts to compensate teachers based on their performance, districts report it is rare for a teacher to not receive the performance bonus. Thus, perhaps the Minnesota Q-Comp program is not truly a high stakes evaluation system if, indeed, all teachers typically receive the performance bonus. Whether tenured teacher Q-Comp evaluations are conducted by peer evaluators or by principals, a reduction in the percentage of staff awarded performance pay might cause the evaluation process to become more high stakes and, perhaps, increase the importance of the *accuracy* practices.

Recommendations

1. It is recommended that principal pre-service and in-service training conducted by principal preparation programs and districts insure that training for principals and principal candidates focuses not only on the technical elements of evaluation but also on the propriety standards of evaluation which address ethical and relationship factors.

2. It is recommended that districts and policy makers review personnel evaluation systems insuring that systems address all aspects of the personnel evaluation standards.

3. It is recommended that superintendent preparation programs address issues of effective evaluation for policy implications given the current political climate of re-designing teacher compensation.

4. It is recommended that districts review their evaluation system policies to insure the accuracy standards are supported as these standards tend to be less practiced by principals.

5. It is recommended that evaluations of principals include components of teacher evaluation.

Suggestions for Further Research

The findings and conclusions in this study lead to the following recommendations for further research and study:

1. Additional studies done based on different populations. Elementary school studies would yield higher sample numbers and possibly a different perspective.

2. A more in-depth study on principals and teachers utilizing qualitative data obtained through focus groups or interviews. Qualitative data would provide in-sights into the complexity of the issues.

3. A more in-depth study that paired teachers with principals. The current study did not match teachers with their supervising principals.

4. Additional study could be done on the effectiveness of peer teacher evaluators as perceived by teachers and peer evaluators.

5. Additional studies could be done involving districts with competitive high stakes performance based evaluation systems.

REFERENCES

Acheson, K. & Gall, M. (1980). *Techniques in the clinical supervision of teachers. Preservice and inservice applications*. New York: Longman, Inc.

Annunziata, J. (1999). Richard Fossey: If a practitioner cleans the windows, will you look in? *Journal of Personnel Evaluation in Education, 13*(1), 83-91.

Atkins, A. O. (1996). Teachers' opinions of the teacher evaluation process**.** East Lansing, MI: National Center for Research on Teacher Learning. (Eric Document Reproduction Service No. ED398628). Retrieved from ERIC (Educational Resources Information Center) database.

Bacharach, S., Conley S., & Shedd, J. (1990). Evaluating teachers for career awards and merit pay. In J. Millman & L. Darling-Hammond (Eds.), *The new handbook of teacher evaluation: Assessing elementary and secondary school teachers* (pp. 133-146). Newbury Park, CA: SAGE Publications Ltd.

Bamberger, M., Rugh, J., Church, M., & Fort, L. (2004). Shoestring evaluation: Designing impact evaluations under budget, time and data constraints. *The American Journal of Evaluation, 25*(1), 5-37.

Barak, M., Pearlman-Avnion, S., & Glanz, J. (1997). Using developmental supervision to improve science and technology instruction in Isreal. *Journal of Curriculum and Supervision, 12*, 367-392.

Barth, M. (2004). A low-cost, post hoc method to rate overall site quality in a multi-site demonstration. *The American Journal of Evaluation, 25*(1), 79-97.

Berk, R. A. (1988). Fifty reasons why student achievement gain does not mean teacher effectiveness. *Journal of Personnel Evaluation in Education, 1*(4), 345-363.

Bernstein, E. (2004). What teacher evaluation should know and be able to do: A commentary. *NASSP Bulletin, 88*, 80-88.

Berube, B., & Dexter, R. (2006). Supervision, evaluation and NCLB: Maintaining a most highly qualified staff. *Catalyst for Change, 34*(2), 11-17.

Blanton, L. P., Sindelar, P.T., & Correa, V.I. (2006). Models and measures of beginning teacher quality. *Journal of Special Education, 40*(2), 115-127.

Bloom, G. (2005). Opt in to an effective supervision process. *Leadership, 34*(5), 34-36.

Bolino, M., & Turnley, W. (2003). Counternormative impression management, likeability, and performance ratings: The use of intimidation in an organizational setting. *Journal of Organizational Behavior, 24*(2), 237-250.

Bommer, W., Johnson, J., Rich, G., Podsakoff, P., & Mackenzie, S. (1995). On the interchangeability of objective and subjective measures of employee performance: A meta-analysis. *Personnel Psychology, 48*(3), 587-605.

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

Bouchamma, Y. (2005). Evaluating teaching personnel. Which model of supervision do Canadian teachers prefer? *Journal of Personnel Evaluation in Education, 18*(4), 289-308.

Brandt, R. (1987a). On teacher evaluation: A conversation with Tom McGreal. *Educational Leadership 44,* 20-24.

Brandt, R. (1987b). Proceed With Caution. *Educational Leadership, 44*(7), 3.

Brandt, R. (1992). On rethinking leadership: A conversation with Tom Sergiovanni. *Educational Leadership, 49*(5), 46-49.

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

Bridges, E. (1990). Evaluation for tenure and dismissal. In J. Millman & L. Darling-Hammond (Eds.), *The new handbook of teacher evaluation: Assessing elementary and secondary school teachers* (pp. 147-157). Newbury Park, CA: SAGE Publications Ltd.

Bridges, E. (1992). *The incompetent teacher: Managerial responses.* Washington, DC: Palmer Press.

Bridges, E., & Groves, B. (1999). The macro- and micropolitics of personnel evaluation: A framework. *Journal of Personnel Evaluation Education, 13*(4), 321-337.

Brundage, S. (1996). What kind of supervision do veteran teachers need? An invitation to expand collegial dialogue and research. *Journal of Curriculum and Supervision, 12,* 90-94.

Bryant, M. (1990). A study of administrative expertise in participant performance on the NASSP assessment center. *Journal of Personnel Evaluation in Education, 3*(4), 353-363.

Buttram, J., & Wilson, B. (1987). Promising trends in teacher evaluation. *Educational Leadership, 44,* 4-6.

Chow, A., Wong, E., Yeung, A., & Mo, K. (2002). Teachers' perceptions of appraiser–appraisee relationships. *Journal of Personnel Evaluation in Education, 16*(2), 85-101.

Clarke, A., & Collins, J. (2004). Glickman's supervisory belief inventory: A cautionary note. *Journal of Curriculum and Supervision, 20*(1), 76-87.

Colby, S., Bradshaw L., & Joyner, R. (April 2002). *Teacher evaluation: A review of the literature.* Paper presented at an annual meeting of the American Educational Research Association, New Orleans, LA. Retrieved from ERIC (Educational Resources Information Center) database.

Conley, D. (1987). Critical Attributes of Effective Evaluation Systems. *Educational Leadership, 44*(7), 60-64.

Conley, S., & Bacharach, S. (1990). Performance appraisal in education: A strategic consideration. *Journal of Personnel Evaluation in Education, 3*(4), 309-319.

Conley, S., Muncey, D., & You, S. (2005). Standards-based evaluation and teacher career satisfaction: A structural equation modeling analysis. *Journal of Personnel Evaluation in Education, 18*(1), 39-65.

Cooper, B., Ehrensal, P., & Bromme, M. (2005). School-level politics and professional development: Traps in evaluating the quality of practicing *teachers. Educational Policy, 19*(1), 112-125.

Copland, M. (2001). The myth of the superprincipal. *Phi Delta Kappan, 82*(7), 528-533.

Costa, A. (1984). Mediating the metacognitive. *Educational Leadership, 42*(3), 57-62.

Costa, A., & Garmston, R. (1994). *Cognitive coaching: A foundation for renaissance schools*. Norwood, MA: Christopher-Gordon Publishers, Inc.

Cruickshank, D., & Haefele, D. (2001). Good teachers, plural. *Educational Leadership, 58*(5), 26-32.

Danielson, C. (1996). Enhancing professional practice: A framework for teaching. Alexandria, VA: ASCD.

Danielson, C. (2001). New trends in teacher evaluation. *Educational Leadership, 58*(5), 12-15.

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

Darling-Hammond, L. (1989). Teacher professionalism and accountability. *Education*

Digest, 55(1), 15.

Darling-Hammond, L., & Snyder, J. (2000). Authentic assessment of teaching in context. *Teaching and Teacher Education, 16*(5-6), 523-545.

Darling-Hammond, L., Wise, A., & Pease, S. (1983). Teacher evaluation in the organizational context: A review of the literature. *Review of Educational Research, 53*(3), 285-328.

Davis, D., Ellett, C., & Annunziata, J. (2002). Teacher evaluation, leadership and learning organizations. *Journal of Personnel Evaluation in Education, 16*(4), 287-301.

Davis, D., Pool, J., & Mits-Cash, M. (2000). Issues in implementing a new teacher assessment system in a large urban school district: Results of a qualitative field study. *Journal of Personnel Evaluation in Education, 14*(4), 285-306.

Davis, S., & Hensley, P. (1999). The politics of principal evaluation. *Journal of Personnel Evaluation in Education, 13*(4), 383-403.

DeSander, M. (2000). Teacher evaluation and merit pay: Legal considerations, practical concerns. *Journal of Personnel Evaluation in Education, 14*(4), 307-17.

Duke, D., & Stiggins, R. (1990). Beyond minimum competence: evaluation for professional development. In J. Millman & L. Darling-Hammond (Eds.), *The new handbook of teacher evaluation: assessing elementary and secondary school teachers* (pp. 116-132). Newbury Park, CA: SAGE Publications Ltd.

Ellett, C., & Garland, J. (1987). Teacher evaluation practices in our largest school districts: Are they measuring up to 'state-of-the-art' systems*? Journal of Personnel Evaluation in Education, 1*(1), 69-92.

Ellett, C., Loup, K., Evans, R., Chauvin, S., & Naik, N. (1994). A study of teachers' nominations of superior colleagues: Implications for teacher evaluation programs and the construct validity of classroom-based assessments of teaching and learning. *Journal of Personnel Evaluation in Education, 8*(1), 7-28.

Ellett, C., Wren, C., Callender, K., Loup, K., & Lui, X. (1996). Looking backwards with the personnel evaluation standards: An analysis of the development and implementation of a statewide teacher assessment program. *Studies in Educational Evaluation, 22*(1), 79-113.

Epstein, J. (1985). A question of merit: Principals' and parents' evaluations of teachers. *Educational Researcher, 14*, 3-10.

Erffmeyer, E., & Martray, C. (1990). A quantified approach to the evaluation of teacher

professional growth and development and professional leadership through a goal-setting process. *Journal of Personnel Evaluation in Education, 3*(3), 275-300.

Follman, J. (1995). Elementary public school pupil rating of teacher effectiveness. *Child Study Journal, 25*(1), 57-78.

Fossey, R. (1998). Secret settlement agreements between school districts and problem employees: Some legal pitfalls. *Journal of Personnel Evaluation in Education, 12*(1), 61-67.

Frase, L., & Streshly, W. (1994). Lack of accuracy, feedback, and commitment in teacher evaluation. *Journal of Personnel Evaluation in Education, 8*(1), 47-57.

Gajda, R. (2004). Utilizing collaboration theory to evaluate strategic alliances. *The American Journal of Evaluation, 25*(1), 65-77.

Gallagher, H. (2004). Vaughn Elementary's innovative teacher evaluation system: Are teacher evaluation scores related to growth in student achievement? *Peabody Journal of Education, 79*(4), 70-107.

George, P. (1987). Performance management in education. *Educational Leadership, 44*(7), 32-39.

Glanz, J. (2005). Action research as instructional supervision: Suggestions for principals. *NASSP Bulletin, 89*(643), 17-27.

Glanz, J., & Neville, R. (Eds.). (1997). *Educational supervision: Perspectives, issues, and controversies.* Norwood, MA: Christopher-Gordon Publishers.

Glass, G. (1990). Using student test scores to evaluate teachers. In J. Millman & L. Darling-Hammond (Eds.), *The new handbook of teacher evaluation: Assessing elementary and secondary school teachers* (pp. 229-240). Newbury Park, CA: SAGE Publications Ltd.

Glatthorn, A., & Holler, R. (1987). Differentiated teacher evaluation. *Educational Leadership, 44*(7), 56-58.

Goe, L. (2007). *The link between teacher quality and student outcomes: A research synthesis.* Washington, DC: National Comprehensive Center for Teacher Quality. Retrieved May 30, 2008, from http://www.ncctq.org/publications/LinkBetweenTQandStudentOutcomes.pdf

Goldstein, J. (2007). Easy to dance to: Solving the problems of teacher evaluation with peer assistance and review. *American Journal of Education, 113*(3), 479-508.

Goldstein, J., & Noguera, P. (2006). A thoughtful approach to teacher evaluation.

Educational Leadership, 63(6), 31-37.

Gordon, S. (2006). Teacher evaluation and professional development. In J. H. Stronge (Ed.), *Evaluating teachers: A guide to current thinking and best practice* (pp. 268-290). Thousand Oaks, CA: Corwin Press.

Hallinger, P., & Heck, R. (1996). Reassessing the principal's role in school effectiveness: A review of empirical research, 1980-1995. *Educational Administration Quarterly, 32*(1), 5-44.

Hallinger, P., & Leithwood, K. (1996). Culture and educational administration. A case of finding out what you don't know you don't know. *Journal of Educational Administration, 34*(5), 18.

Halverson, R., & Clifford, M. (2006). Evaluation in the wild: A distributed cognition perspective on teacher assessment. *Educational Administration Quarterly, 42*(4), 578-619.

Harris, B. (1987). Resolving old dilemmas in diagnostic evaluation. *Educational Leadership, 44*(7), 46-50.

Heath, R., & Nielson, M. (1974). The research basis for performance-based teacher education. *Review of Educational Research, 44*(4), 463-484.

Heneman, H., & Milanowski, A. (1999). Teachers' attitudes about teacher bonuses under school-based performance award programs. *Journal of Personnel Evaluation in Education, 12*(4), 327-341.

Heneman, H., & Milanowski, A. (2003). Continuing assessment of teacher reactions to a standards-based teacher evaluation system. *Journal of Personnel Evaluation in Education, 17*(2), 173-195.

Heneman, H., Milanowski, A., Kimball, S., & Odden, A. (2006). *Standards-based teacher evaluation as a foundation for knowledge- and skill-based pay.* Philadelphia: Consortium for Policy Research in Education.

Heneman, R. (1986). The relationship between supervisory ratings and results-oriented measures of performance: A meta-analysis. *Personnel Psychology, 39*(4), 811-826.

Holland, P. (2004). Principals as supervisors: A balancing act. *NASSP Bulletin, 88*(639), 3-14.

Holland, P., & Adams, P. (2002). Through the horns of a dilemma between instructional supervision and the summative evaluation of teaching. *International Journal of Leadership in Education, 5*(3), 227-247.

Holland, P., & Garman, N. (2001). Toward a resolution of the crisis of legitimacy in the field of supervision. *Journal of Curriculum and Supervision, 16*(2), 95-112.

Howard, B. (2005). *Teacher growth and assessment process procedural handbook.* University of North Carolina, Greensboro, NC: SERVE Center. (ERIC Document Reproduction Service No. ED485281) Retrieved on April 11, 2008 from ERIC (Educational Resources Information Center) database.

Howard, B., & McColskey, W. (2001). Evaluating experienced teachers. *Educational Leadership, 58*(5), 48-52.

Howard, B., & Sanders, J. (2006). Applying the personnel evaluation standards to teacher evaluation. In J. Stronge, (Ed.), *Evaluating Teaching: A Guide to Current Thinking and Best Practice* (pp. 54-68). Thousand Oaks, CA: Corwin Press**.**

Hoxie, C., & Leigh, A. (2004). Pulled away or pushed out? Explaining the decline of teacher aptitude in the United States. *The American Economic Review, 94*(2), 236-240.

Imig, D., & Imig, S. (2006). The teacher effectiveness movement: How 80 years of essentialist control have shaped the teacher education profession. *Journal of Teacher Education, 57*(2), 167-180.

Iwanicki, E. (1990). Teacher evaluation for school improvement. In J. Millman & L. Darling-Hammond (Eds.), *The new handbook of teacher evaluation: Assessing elementary and secondary school teachers* (pp. 147-157). Newbury Park, CA: SAGE Publications.

Jacob, B. (2005). Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics, 89*(5-6), 761-796.

Jacob, B., & Lefgren, L. (2005). Principals as agents: Subjective performance measurement in education. National Bureau of Economic Research Working Paper Series No. 11463. Retrieved April 11, 2008, from http://www.nber.org/papers/w11463.pdf

Jacob, B., & Lefgen, L. (2006). When principals rate teachers: The best--and the worst--stand out. *Education Next, 2,* 58-64.

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

Jacob, B., & Levitt, S. (2003). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *Quarterly Journal of Economics, 118*(3), 843-877.

Johnson, S. (1992). *Teachers at work: Achieving success in our schools.* New York: Basic Books.

Joint Committee on Standards for Educational Evaluation. (1981). *Standards for evaluations of educational programs, projects, and materials.* (D. Stufflebeam, chair). New York: McGraw-Hill.

Joint Committee on Standards for Educational Evaluation. (1988). *The personnel evaluation standards: How to assess systems for evaluating educators.* (D. Stufflebeam, chair). Newbury Park, CA: SAGE Publications.

Joint Committee on Standards for Educational Evaluation. (2009). *The personnel evaluation standards: How to assess systems for evaluating educators.* (2nd ed.). (A. Gullickson, chair). Thousand Oaks, CA: Corwin Press.

Kauchak, D., Peterson, D., & Driscoll, A. (1985). An interview study of teachers' attitudes toward teacher evaluation practices. *Journal of Research and Development in Education, 19*(1), 5.

Kerrins, J., & Cushing, K. (2000). Taking a second look: Expert and novice differences when observing the same classroom teaching segment a second time. *Journal of Personnel Evaluation in Education, 14*(1), 5-24.

Kersten, T., & Israel, M. (2005). Teacher evaluation: Principals' insights and suggestions for improvement. *Planning and Changing, 36*(1), 47-67.

Kiley, M. (1988). *Teachers' and administrators' views of evaluation--differing perspectives.* (ERIC Document Reproduction Service No. ED300434) Retrieved on April 20, 2008 from ERIC (Educational Resources Information Center) database.

Kimball, S. (2002). Analysis of feedback, enabling conditions and fairness perceptions of teachers in three school districts with new standards-based evaluation systems. *Journal of Personnel Evaluation in Education, 16*(4), 241-68.

Kimball, S., White, B., Milanowski, A., & Borman, G. (2004). Examining the relationship between teacher evaluation and student assessment results in Washoe County. *Peabody Journal of Education, 79*(4), 54-78.

Kyriakides, L., & Demetriou, D. (2007). Introducing a teacher evaluation system based on teacher effectiveness research: An investigation of stakeholders' perceptions. *Journal of Personnel Evaluation in Education, 20*(1), 43-64.

Kyriakides, L., Demetriou, D., &Charalambous, C. (2006). Generating criteria for evaluating teachers through teacher effectiveness research. *Educational Research, 48*(1), 1-20.

Ladd, H. (1997). The Dallas school accountability and incentive program: An evaluation of its impacts on student outcomes. *Economics of Education Review, 18*(1999), 1-16.

Little, J. (1982). Norms of collegiality and experimentation: Workplace conditions of school success. *American Educational Research Journal, 19*(3), 325-340.

Loup, K., & Ellett, C. (1997). Application of the personnel evaluation standards to local district teacher evaluation programs: Analyses of 14 cases. (ERIC Document Reproduction Service No. ED412224) Retrieved on April 20, 2008 from ERIC (Educational Resources Information Center) database.

Loup, K., Garland, J., Ellett, C., & Rugutt, J. (1996). Ten years later: Findings from a replication of a study of teacher evaluation practices in our 100 largest school districts. *Journal of Personnel Evaluation in Education, 10*(3), 203-226.

Manatt, R. (1982). *The School Improvement Model: A scenario for operational status, 1983-1984*. St. Paul, MN: Iowa State University of Science and Technology, Ames. Research Institute for Studies in Education, Northwest Area Foundation, St. Paul. (ERIC Document Reproduction Service No. ED225278) Retrieved on April 20, 2008 from ERIC (Educational Resources Information Center) database.

Manatt, R. (1987). Lessons from a comprehensive performance appraisal project. *Educational Leadership, 44*(7), 8-15.

Manatt, R. (1988). Teacher performance evaluation: A total system approach. In S. Stanley & W. Popham (Eds.). *Teacher evaluation: Six prescriptions for success* (pp. 79-108). Alexandria, VA: ASCD.

Manatt, R. (1994 July). *A total systems approach to performance evaluation: How the school improvement model (SIM) uses evaluation to improve teaching and learning.* Paper presented at the Annual National Evaluation Institute of the Center for Research on Educational Accountability and Teacher Evaluation, Gatlinburg, TN.

Manatt, R., & Daniels, B. (1990). Relationships between principals' ratings of teacher performance and student achievement. *Journal of Personnel Evaluation in Education, 4*, 189-191.

Marshall, K. (2005). It's time to rethink teacher supervision and evaluation. *Phi Delta Kappan, 86*(10), 727.

Mathers, C., Oivia, M., & Laine, S. (2008*). Improving instruction for effective teacher evaluation: Options for states and districts.* TQ Teacher Quality and Policy Brief. Washington, DC: National Comprehensive Center for Teacher Quality**.**

McCaffrey, D., Lockwood, J., Koretz, D., Louis, T., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics, 29*(1), 67-102.

McGrath, M. J. (1995). Effective evaluation. *Thrust for Educational Leadership, 24*(6), 36-39.

McIntire, R., Hughes, L., & Burry, J. (1987). The training and certifying of teacher appraisers. *Educational Leadership, 44*(7), 62-65.

McLaughlin, M. (1988). *Teacher evaluation: improvement, accountability, and effective learning.* New York: Teachers College Press.

McLaughlin, M. (1990). Embracing contraries: Implementing and sustaining teacher evaluation. In J. Millman & L. Darling-Hammond (Eds.). *The new handbook of teacher evaluation: Assessing elementary and secondary school teachers* (pp. 403-415). Newbury Park, CA: SAGE Publications.

Medley, D., & Coker, H. (1987). The accuracy of principals' judgments of teacher performance. *Journal of Educational Research, 89*(4), 242-247.

Mehrens, W. (1990). Combining evaluation data from several sources. In J. Millman & L. Darling-Hammond (Eds.). *The new handbook of teacher evaluation: Assessment of elementary and secondary school teachers* (pp. 322-336). Newbury Park, CA: SAGE Publications.

Mendro, R. (1998). Student achievement and school and teacher accountability. *Journal of Personnel Evaluation in Education, 12*, 257-267.

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

Milanowski, A., & Heneman, H. (2001). Assessment of teacher reactions to a standards-based teacher evaluation system: A pilot study. *Journal of Personnel Evaluation in Education, 15*(3), 193-212.

Millman, J. (1981). Student achievement as a measure of teacher competence. In J. Millman (Ed.), *Handbook of teacher evaluation* (pp. 156-166) Beverly Hills, CA: SAGE Publications.

Murphy, J. (1987). Teacher evaluation: A comprehensive framework for supervisors. *Journal of Personnel Evaluation in Education, 1*(2), 157-180.

National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform. An open letter to the American people. A*

report to the nation and to the secretary of education. (D. Gardner, chair). Washington, D.C.: Department of Education. (ERIC Document Reproduction Service No. ED226006) Retrieved on April 11, 2008 from ERIC (Educational Resources Information Center) database.

Natriello, G. (1990). Intended and unintended consequences: Purposes and effects of teacher evaluation. In J. Millman & L. Darling-Hammond (Eds.), *The new handbook of teacher evaluation: Assessing elementary and secondary school teachers* (pp. 35-45). Newbury Park, CA: SAGE Publications.

Nye, B., Konstantopoulos, S., & Hedges, L. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis, 26*(3), 237-257.

Odden, A. (2004). Lessons learned about standards-based teacher evaluation systems. *Peabody Journal of Education, 79*(4), 11.

Ostrander, L. (1996). Multiple judges of teacher effectiveness: Comparing teacher self-assessments with the perceptions of principals, students, and parents. (ERIC Document Reproduction Service No. ED399267) Retrieved on April 21, 2008 from ERIC (Educational Resources Information Center) database.

Ovando, M. (2005). Building instructional leaders' capacity to deliver constructive feedback to teachers. *Journal of Personnel Evaluation in Education, 18*(3), 171-183.

Painter, S. (2000). Principals' efficacy beliefs about teacher evaluation. *Journal of Educational Administration, 38*(4), 368-378.

Peterson, K. (1987). Teacher evaluation with multiple and variable lines of evidence. *American Educational Research Journal, 24*(2), 311-317.

Peterson, K. (2000). *Teacher evaluation: A comprehensive guide to new directions and practices.* Thousand Oaks, CA: Corwin Press.

Peterson, K. (2004). Research on school teacher evaluation. *NASSP Bulletin, 88*(639), 60-79.

Peterson, K., & Chenoweth, T. (1992). School teachers' control and involvement in their own evaluation. *Journal of Personnel Evaluation in Education, 6*(2), 177-90.

Peterson, K., Kelley, P., & Caskey, M. (2002). Ethical considerations for teachers in the evaluation of other teachers. *Journal of Personnel Evaluation in Education, 16*(4), 317-324.

Peterson, K., & Peterson, C. (2005). *Effective teacher evaluation: A guide for principals*. Thousand Oaks, CA: Corwin Press.

Peterson, K., Stevens, D., & Mack, C. (2001). Presenting complex teacher evaluation data: Advantages of dossier organization techniques over portfolios. *Journal of Personnel Evaluation in Education, 15*(2), 121-133.

Peterson, K., Wahlquist, C., & Bone, K. (2000). Student surveys for school teacher evaluation. *Journal of Personnel Evaluation in Education, 14*(2), 135-153.

Peterson, K., Wahlquist, C., Brown, J., & Mukhopadhyay, S. (2003). Parent surveys for teacher evaluation. *Journal of Personnel Evaluation in Education, 17*(4), 317-330.

Podgursky, M., Monroe, R., & Watson, D. (2004). The academic quality of public school teachers: An analysis of entry and exit behavior. *Economics of Education Review, 23*(5), 507-518.

Podgursky, M., & Springer, M. (2007). Teacher performance pay: A review. *Journal of Policy Analysis and Management, 26*(4), 909-949.

Ponticell, J., & Zepeda S. (2004). Confronting well-learned lessons in supervision and evaluation. *NASSP Bulletin, 88*(639), 43-59.

Pool, J., Ellett, C., Schiavone, S., & Carey-Lewis, C. (2001). How valid are the national board of professional teaching standards assessments for predicting the quality of actual classroom teaching and learning? Results of six mini case studies. *Journal of Personnel Evaluation in Education, 15*(1), 31-48.

Popham, W. (1987). The shortcomings of champagne teacher evaluations. *Journal of Personnel Evaluation in Education, 1*(1), 25-28.

Popham, W. (1988a). The dysfunctional marriage of formative and summative teacher evaluation. *Journal of Personnel Evaluation Education, 1,* 269-673.

Popham, W. (1988b). Judgment-based teacher evaluation. In W. Popham & S. Stanley (Eds.), *Teacher evaluation: Six prescriptions for success* (pp. 56-78). Alexandria, VA: ASCD.

Reitzug, U. (1997). Images of principal instructional leadership: From super-vision to collaborative inquiry. *Journal of Curriculum and Supervision, 12*(4), 324-343.

Renger, R., & Bourdeau, B. (2004). Strategies for values inquiry: An exploratory case study. *The American Journal of Evaluation, 25*(1), 39-49.

Ribas, W. (2005). *Teacher evaluation that works!!* Westwood, MA: Ribas Publications.

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

Rogers, P. (2004). 2002 AEA awards. *The American Journal of Evaluation, 25*(1), 129-135.

Rooney, J. (2005). Teacher supervision: If it ain't working. *Educational Leadership, 63*(3), 88-89.

Sanders, J., Saxton, A., & Horn, S. (1997). The Tennessee value-added assessment system: A quantitative, outcomes-based approach to educational assessment. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid evaluation measure?* (pp. 137-162). Thousand Oaks, CA: Corwin Press.

Sanders, W., & Rivers, J. (1996). Cumulative and residual effects of teachers on future student academic achievement. University of Tennessee Value-Added Research and Assessment Center. Knoxville, TN: University of Tennessee**.**

Sanders, W., & Horn, S. (1998). Research findings from the Tennessee value-added assessment system (TVAAS) database: Implications for educational evaluation and research. *Journal of Personnel Evaluation in Education, 12*(3), 247-256.

Sanders, W., Wright, S., & Horn, S. (1997). Teacher and classroom context effects on student achievement: Implications for teacher evaluation. *Journal of Personnel Evaluation in Education, 11*(1), 57-67.

Schon, D. (1983). *The reflective practitioner: How professionals think in action.* New York: Basic Books

Schon, D. (1987). *Educating the reflective practitioner: Toward a new design for teaching and learning in the professions.* San Francisco, CA: Jossey-Bass.

Scriven, M. (1967). The methodology of evaluation. In R. Tyler & M. Scriven (Eds.), *AERA monograph review on curriculum evaluation: No. 1.* (pp. 39-83). Chicago, IL: Rand McNally**.**

Scriven, M. (1981). Summative teacher evaluation. In J. Millman (Ed.), *Handbook of teacher evaluation* (pp. 244-271). Beverly Hills, CA: SAGE Publications.

Scriven, M. (1987). Validity in personnel evaluation. *Journal of Personnel Evaluation in Education, 1*(1), 9-23.

Scriven, M. (1988). Duty-based teacher evaluation. *Journal of Personnel Evaluation in Education, 1*, 310-332.

Scriven, M. (1988). Evaluating teachers as professionals: The duties-based approach. In S. Stanley & W. Popham (Eds.), *Teacher Evaluations: Six Prescriptions for Success* (pp. 100-143). Alexandrea, VA: ASCD**:** 100-143.

Scriven, M. (1990). Can research-based teacher evaluation be saved? *Journal of Personnel Evaluation in Education, 4*(1), 19-32.

Scriven, M. (1994). Duties of the teacher. *Journal of Personnel Evaluation in Education, 8*(2), 151-184.

Scriven, M. (1995a). Student ratings offer useful input to teacher evaluations. *ERIC Digests*. Washington, DC**:** ERIC Clearinghouse on Assessment and Evaluation. (ERIC Document Reproduction Service No. ED398240) Retrieved on April 21, 2008 from ERIC (Educational Resources Information Center) database.

Scriven, M. (1995b). A unified theory approach to teacher evaluation. *Studies In Educational Evaluation, 21*(2), 111-129.

Scriven, M. (2002). Out of the frying pan, into the fire: Comments on Roth/Tobin. *Journal of Personnel Evaluation in Education, 16*(4), 303-06.

Shinkfield, A., & Stufflebeam, D. (1995). *Teacher evaluation: guide to effective practice*. Norwell, MA: Kluwer Academic Publishers.

Showers, B. (1984). *Peer coaching: A strategy for facilitating transfer of training.* Oregon University, Eugene, OR: Center for Educational Policy and Management**.**

Siens, C., & Ebmeier, H. (1996). Developmental supervision and the reflective thinking of teachers*. Journal of Curriculum and Supervision, 11*(4), 299-319.

Smylie, M. (1996). From bureaucratic control to building human capital: The importance of teacher learning in education reform. *Educational Researcher, 25*(9), 9-11.

Spicer, R., Nelkin, V., Miller, T., & Becker, L. (2004). Using corporate data in workplace program evaluation. *The American Journal of Evaluation, 25*(1), 109-119.

Stake, B. (2004). How far dare an evaluator go toward saving the world? *The American Journal of Evaluation, 25* (1) 103-107.

Starratt, R. (1992). After supervision. *Journal of Curriculum and Supervision, 8*, 9.

Sternberg, R., & Horvath, J. (1995). A prototype view of expert teaching. *Educational Researcher, 24*(6), 9-17.

Stiggins, R. (1986). Teacher evaluation: Accountability and growth systems-different purposes. *NASSP Bulletin, 70*, 51-59.

Stiggins, R. (1989). A commentary on the role of student achievement data in the evaluation of teachers. *Journal of Personnel Evaluation in Education, 3*, 7-15.

Stiggins, R., & Bridgeford, N. (1984). *Performance assessment for teacher development.* (ERIC Document Reproduction Service No. ED242717) Retrieved on April 28, 2008 from ERIC (Educational Resources Information Center) database.

Stiggins, R., & Duke, D. (1988). *The case for commitment to teacher growth.* Albany, NY: State University of New York Press.

Stodolsky, S. (1984). Teacher evaluation: The limits of looking. *Educational Researcher 13*, 11-18.

Stodolsky, S. (1990). Classroom observation. In J. Millman & L. Darling-Hammond (Eds.), *The new handbook of teacher evaluation: Assessing elementary and secondary school teachers* (pp. 175-190). Newbury Park, CA: SAGE Publications.

Stronge, J. (1995). Balancing individual and institutional goals in educational personnel evaluation: A conceptual framework. *Studies in Educational Evaluation, 21*(2), 131-151.

Stronge, J. (2002). *Qualities of effective teachers,* (ERIC Document Reproduction Service No. ED468204) Retrieved on April 21, 2008 from ERIC (Educational Resources Information Center) database.

Stronge, J. (2006). Teacher evaluation and school improvement: Improving the educational landscape. In J. Stronge (Ed.), *Evaluating teaching: A guide to current thinking and best practice* (pp. 1-23). Thousand Oaks, CA: SAGE Publications.

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

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

Stufflebeam, D. (1997). Oregon teacher work sample methodology: Educational policy review. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid evaluation measure?* (pp. 53-61). Thousand Oaks, CA: Corwin Press.

Stufflebeam, D. (2004). A note on the purposes, development, and applicability of the Joint Committee Evaluation Standards. *The American Journal of Evaluation, 25*(1), 99-102.

Stufflebeam, D., & Brethower, D. (1987). Improving personnel evaluations through professional standards. *Journal of Personnel Evaluation Education, 1,* 125-155.

Stufflebeam, D., & Pullin, D. (1998). Achieving legal viability in personnel evaluations. *Journal of Personnel Evaluation in Education, 11*(3), 215-230.

Stufflebeam, D., & Sanders, J. (1990). Using the Personnel Evaluation Standards to improve teacher evaluation. In J. Millman & L. Darling-Hammond (Eds.), *The new handbook of teacher evaluation: Assessing elementary and secondary school teachers* (pp. 416-428). Newbury Park, CA: SAGE Publications**.**

Tell, C. (2001). Appreciating good teaching. *Educational Leadership, 58*(5), 6-12.

Tesch, S., Nyland, L., & Kernutt, D. (1987). Teacher evaluation--shared power working. *Educational Leadership, 44*(7), 26-31.

Thorson, J., Miller, R., & Bellon, J. (1987). Instructional improvement through personnel evaluation. *Educational Leadership, 44*(7), 52-55.

Timperley, H. (1998). Performance appraisal: Principals' perspectives and some implications. *Journal of Educational Administration, 36*(1), 44-58.

Tobin, K., & Roth, W. (2002). Concerning the fallibility of judgments from the side, the rear, and on high: A dialogue about Scriven's critique. *Journal of Personnel Evaluation in Education, 16*(4), 307-314.

Torff, B., & Sessions, D. (2005). Principals' perceptions of the causes of teacher ineffectiveness. *Journal of Educational Psychology, 97*(4), 530-537.

Tucker, P. (1997). Lake Wobegon: Where all teachers are competent (Or, have we come to terms with the problem of incompetent teachers?). *Journal of Personnel Evaluation in Education, 11*(2), 103-126.

Tucker, P. (2001). Helping struggling teachers. *Educational Leadership, 58*(5), 52-59.

Tucker, P., & Stronge, J. (2006). Student achievement and teacher evaluation. In J. Stronge (Ed.), *Evaluating teaching: A guide to current thinking and best practice* (pp. 152-167). Thousand Oaks, CA: Corwin Press.

Turner, G., & Clift, P. (1988). *Studies in teacher appraisal: a project funded by the Leverhulme Trust.* Philadelphia, PA: The Falmer Press.

Waite, D. (1996). Sociocultural research questions for supervision. *Journal of Curriculum and Supervision, 11*(3), 289-294.

Wang, W., & Day, C. (2001 February). *Issues and concerns about classroom observation: Teachers' perspectives.* Paper presented at the Annual Meeting of the Teachers of English to Speakers of Other Languages (TESOL), St. Louis, MO.

Weston, T. (2004). Formative evaluation for implementation: Evaluating educational technology applications and lessons. *The American Journal of Evaluation, 25*(1), 51-64.

Wheeler, P., & Scriven, S. (2006). Building the foundation: Teacher roles and responsibilities. In J. Stronge (Ed.), *Evaluating teaching: A guide to current thinking and best practice* (pp. 25-53). Thousand Oaks, CA: Corwin Press.

Wilkerson, D., Manatt, R., Rogers, M., & Maughan, R. (2000). Validation of student, principal, and self-ratings in 360° feedback® for teacher evaluation. *Journal of Personnel Evaluation in Education, 14*(2), 179-192.

Wise, A., Darling-Hammond, L., McLaughlin, M., & Berstein, H. (1984a). *Case studies for teacher evaluation: A study of effective practices.* Santa Monica, CA: The Rand Corporation**.**

Wise, A., Darling-Hammond, L., McLaughlin, M., & Bernstein, H. (1984b). *Teacher evaluation: A study of effective practices.* Santa Monica, CA: The Rand Corporation**:** 101.

Wolf, K. (2006). Portfolios in teacher evaluation. In J. Stronge (Ed.), *Evaluating teaching: A guide to current thinking and best practices* (pp. 168-185). Thousand Oaks, CA: Corwin Press**.**

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

Zepeda, S. (2006). High stakes supervision: we must do more. *International Journal of Leadership in Education, 9*(1), 61-73.

Zepeda, S. (2007). *The principal as instructional leader: A handbook for supervisors.* Larchmont, NY: Eye On Education.

Zimmerman, S., & Deckert-Pelton, M. (2003). Evaluating the evaluators: Teachers' perceptions of the principal's role in professional evaluation. *NASSP Bulletin, 87*(636), 28-37.

APPENDIX A

LIST OF QUALIFYING SCHOOLS

Middle Schools Participating in Q-Comp Program in 2007-2008

	District	School	Dedicated Administrator	Affects Compensation	Annual Eval
1	Alexandria	Discovery Middle School	Yes	No	No
2	Brainerd	Forestview Middle School 5-6	Yes	No	N/A
3	Brainerd	Forestview Middle School 7-8	Yes	No	N/A
4	Brooklyn Center	Brooklyn Center Junior	No	N/A	N/A
5	Burnsville	Eagle Ridge Junior High	Yes	Yes	No
6	Burnsville	Metcalf Junior High School	Yes	Yes	No
7	Burnsville	Nicollet Junior High School	Yes	Yes	No
8	Farmington	Farmington Middle School East	Yes	Yes	No
9	Farmington	Farmington Middle School West	Yes	Yes	No
10	Forest Lake	Century Junior High	Yes	Yes	No
11	Forest Lake	Southwest Junior High	Yes	Yes	No
12	**Fridley**	**Middle School**	**Yes**	**Yes**	**Yes**
13	Grand Meadow	Grand Meadow Middle School	No	N/A	N/A
14	Hopkins	West Junior High School	Yes	No	N/A
15	Hopkins	North Junior High School	Yes	No	N/A
16	Lac Qui Parle	LqPV Middle School	Yes	Yes	No
17	LaCrescent-Hokah	La Crescent Middle School	No	N/A	N/A
18	Le Center	LeCenter Middle School	No	N/A	N/A
19	**Marshall**	**Marshal Middle School**	**Yes**	**Yes**	**Yes**
20	Minneapolis	Anwatin Middle School	Yes	Yes	No
21	Minneapolis	Anthony Middle School	Yes	Yes	No
22	Minneapolis	Folwell Middle School	Yes	Yes	No
23	Minneapolis	Northeast Middle School	Yes	Yes	No
24	Minneapolis	Olson Middle School	Yes	Yes	No
25	Minneapolis	Sanford Middle School	Yes	Yes	No
26	Minnetonka	Minnetonka Middle School East	Yes	Yes	No
27	Minnetonka	Minnetonka Middle School West	Yes	Yes	No
28	North St. Paul-Maplewood-Oakdale	John Glen Middle School	Yes	No	N/A
29	North St. Paul-Maplewood-Oakdale	Maplewood Middle School	Yes	No	N/A
30	North St. Paul-Maplewood-Oakdale	Skyview Middle School	Yes	No	N/A
31	Osseo	Brooklyn Junior	Yes	No	N/A
32	Osseo	Maple Grove Junior	Yes	No	N/A
33	Osseo	North View Junior	Yes	No	N/A

34	Osseo	Osseo Junior	Yes	No	N/A
35	**Proctor**	**Jedlicka Middle School**	**Yes**	**Yes**	**Yes**
36	South Washington County	Lake Junior High School	Yes	Yes	No
37	South Washington County	Cottage Grove Junior High	Yes	Yes	No
38	South Washington County	Oltman Junior High	Yes	Yes	No
39	South Washington County	Woodbury Junior High	Yes	Yes	No
40	**St. Anthony-New Brighton**	**St. Anthony Middle School**	**Yes**	**Yes**	**Yes**
41	**St. Cloud**	**St. Cloud North Junior High**	**Yes**	**Yes**	**Yes**
42	**St. Cloud**	**St. Cloud South Junior High**	**Yes**	**Yes**	**Yes**
43	St. Francis	St. Francis Middle School	Yes	No	N/A

Note: Bolded schools met the requirements to participate in the study.

APPENDIX B

SURVEYS

Teacher Survey

Please take time to consider which principal in your career was the most competent in the area of teacher evaluation. Next, read each item carefully and indicate the degree to which that principal's evaluation practices are like the statements listed below by clicking on the down arrow and selecting a response. Next, indicate how important the item is as a component of an effective evaluation. If a specific item is not applicable to your situation, leave unanswered.

Definition: Evaluatees are the staff being evaluated by the principal

1. Interactions with Evaluatee (Teacher)
The principal…..

Frequency*
Importance*

a) Conducted the evaluation in a respectful manner.
b) Conducted the evaluation in a manner that I considered fair.
c) Communicated the results clearly and objectively.
d) Developed a relationship of mutual trust and understanding with me prior to the evaluation.
e) Demonstrated a genuine interest in me as a person.
f) Took into account my personal and professional needs.

2. Comprehensive Evaluation
The principal….

Frequency*
Importance*

a) Ensured I knew what would be assessed.
b) Ensured I knew how the evaluation data would be collected.
c) Described and justified the basis for interpretation of both positive and negative assessment information and results.
d) Reported fully both strengths and weaknesses with supporting evidence.
e) Supported continued, timely professional growth.

3. Explicit Criteria
The principal…

Frequency*
Importance*

a) Addressed only identified professional roles and responsibilities in the evaluation report and ensured that extraneous comments beyond the criteria were neither included nor accepted.
b) Provided copies of written evaluation reports to me.
c) Used the agreed upon criteria, providing a rationale and justification of evaluation findings.

4. Procedures
The principal…

Frequency*
Importance*

a) Delineated the procedures by which I could exercise my rights to review data about my performance.
b) Encouraged me to suggest ways by which evaluation procedures could be made more efficient and useful.

5. Interpretation
The principal…

Frequency*
Importance*

a) Considered my background and cultural experiences when

interpreting performance.

b) Ensured that my scoring was not influenced by factors irrelevant to my performance being evaluated (general impression or previous rating influences the present rating).
c) Ensured that my summary conclusions which were derived from a series of assessments corresponded with the documented results.

6. Bias Control
The principal...

Frequency*
Importance*

a) Obtained data and judgments from multiple sources to ensure validity and consistent indications of my performance.
b) Allowed me to review data and participate in interpreting it where appropriate.
c) Based my evaluations on defensible information with conclusions that were justifiable.

7. Justified Conclusions
The principal...

Frequency*
Importance*

a) Generated, assessed, and reported plausible alternative explanations of findings and, if appropriate, indicated why these explanations should be discounted.
b) Limited conclusions to those situations, time periods, contexts, and purposes for which the evaluation findings were applicable.

8. What is your gender?

- ☐ Male
- ☐ Female

9. What is your age?

- ☐ 29 or under
- ☐ 30-34
- ☐ 35-39
- ☐ 40-44
- ☐ 45-49
- ☐ 50-54
- ☐ 55-59
- ☐ 60 or older

10. What is your educational background?

- ☐ B.A. or B.S.
- ☐ M.A. or Ed.M.
- ☐ Educational Specialist or 6th Year Certification
- ☐ Ed.D. or Ph.D.
- ☐ Other (please specify)

Note: Frequency response options were: Never, Rarely, Occasionally, Often, Always

Note: Importance response options were: Not Important, Somewhat Important, Important

Principal Survey

Below is a list of descriptors of teacher evaluation practices. Please read each item carefully and indicate the degree to which you perceive or have been told your evaluation practices reflect the descriptor. If a specific item is not applicable to your situation, leave unanswered.

Definition: Evaluatees are the staff being evaluated by the principal

1. Interactions with Evaluatees (Teachers)

		Never	Rarely	Occasionally	Often	Always
a)	I conduct evaluations in a respectful manner.					
b)	I conduct evaluations in such a manner that it is considered fair.					
c)	I communicate results clearly and objectively.					
d)	I develop a relationship of mutual trust and understanding with the teacher prior to the evaluation.					
e)	I demonstrate a genuine interest in the teacher as a person.					
f)	I take into account the teacher's personal and professional needs.					

2. Comprehensive Evaluation

		Never	Rarely	Occasionally	Often	Always
a)	I ensure teachers know what will be assessed.					
b)	I ensure the teachers know how evaluation data will be collected.					
c)	I describe and justify the basis for interpretation of both positive and negative assessment information and results.					
d)	I report fully both strengths and weaknesses and include supporting evidence.					
e)	I support continued and timely professional growth.					

3. Explicit Criteria

		Never	Rarely	Occasionally	Often	Always
a)	I address only the teachers' identified professional roles and responsibilities in the evaluation report.					
b)	I provided copies of written evaluation reports to me.					
c)	I used the agreed upon criteria, providing a rationale and justification of evaluation findings.					

4. Procedures

		Never	Rarely	Occasionally	Often	Always
a)	I delineated the procedures by which teachers can exercise their rights to review data about their performance.					
b)	I encourage teachers to suggest ways by which evaluation procedures can be made more efficient and useful.					

5. Interpretation

		Never	Rarely	Occasionally	Often	Always
a)	I consider the background and cultural experiences of teachers when interpreting performance.					
b)	I ensured that scoring is not influenced by factors irrelevant to the performance being evaluated (general tendency to be too generous or too severe; general impression or previous rating influences the present rating; teaching style attributes).					
c)	I ensure that summary conclusions derived from a series of assessments correspond with the original documented results.					

6. Bias Control

		Never	Rarely	Occasionally	Often	Always
a)	I obtain data and judgments from multiple sources to ensure validity and consistent indications of performance.					
b)	I allow teachers to review data and participate in interpreting					

it where appropriate.

c) I based evaluations on defensible information with conclusions that are justifiable.

7. Justified Conclusions

	Never	Rarely	Occasionally	Often	Always
a) I generate, assess, and report plausible alternative explanations of findings and, if appropriate, indicate why these explanations should be discounted.					
b) I limit conclusions to those situations, time periods, contexts, and purposes for which the evaluation findings are applicable.					

8. What is your gender?

- Male
- Female

9. What is your age?

- 29 or under
- 30-34
- 35-39
- 40-44
- 45-49
- 50-54
- 55-59
- 60 or older

10. How many years have you been a principal?

- 1 – 5
- 6 – 10
- 11 - 15
- More than 15
- Other (please specify)

APPENDIX C

CONSENT FORMS

Letter to District Superintendents

Dear Superintendent:

As part of the research for my doctoral degree in Educational Policy and Administration at the University of Minnesota, I am conducting a survey on the effectiveness of principals as evaluators of teachers for purposes of awarding merit pay. The study focuses specifically on those middle schools in Minnesota in which the principal's evaluation affects the awarding of merit pay. As the purpose of this study is to aggregate views from the principal and teacher perspective, no individual school, principal, or teacher is being tracked, linked, or identified. The study has minimal risks.

As the number of schools meeting the parameter is limited in number, I would appreciate your support of this research project.

Participation will involve asking the principal that evaluates the middle school teachers and the tenured middle school teachers to complete a 10-20 minute online survey. A preview of the principals' questions being asked can be seen here: http://www.surveymonkey.com/s.aspx?sm=LDsAvv7HsV8CdvfW9Yx6Ig_3d_3d. A preview of the teachers' questions can be seen here: http://www.surveymonkey.com/s.aspx?sm=bqznzfw5IFRXRlft_2fH0SxA_3d_3d. Printed copies of the on-line surveys are also attached. All responses will remain anonymous.

Benefits for participation in this survey include opportunities to share information regarding principal evaluation practices. Your assistance in the survey will add to the limited research in this area. The aggregated results of this study will be used to inform principal training organizations. Aggregated results will also be shared with the Minnesota Department of Education, Minnesota Association of Secondary School Principals, and principal training colleges and universities. Furthermore, at the conclusion of the study, all participants will have the opportunity to view the survey results on-line.

If you have any questions or concerns regarding this study and would like to talk to someone other than the researcher, you are encouraged to contact the Research Subjects' Advocate Line, D528 Mayo, 420 Delaware St. Southeast, Minneapolis, Minnesota 55455; 612-625-1650. You can contact me at [address and phone number inserted].

Thank you for your time and assistance.

Sincerely,

Shirley Gregoire

University of Minnesota, Doctoral Candidate
Principal, St. Anthony Middle School in St. Anthony, MN

Letter to Principals

Dear Principal:

As part of the research for my doctoral degree in Educational Policy and Administration at the University of Minnesota, I am conducting a survey on the effectiveness of principals as evaluators of teachers for purposes of awarding merit pay. The study focuses specifically on those middle schools in Minnesota in which the principal's evaluation affects the awarding of merit pay. As the purpose of this study is to aggregate views from the principal and teacher perspective, no individual school, principal, or teacher is being tracked, linked, or identified. The study has minimal risks.

As the number of schools meeting the parameter is limited in number, I would appreciate your support of this research project.

Participation will involve asking the principal that evaluates the middle school teachers and the tenured middle school teachers to complete a 10-20 minute online survey. A preview of the principals' questions being asked can be seen here: http://www.surveymonkey.com/s.aspx?sm=LDsAvv7HsV8CdvfW9Yx6Ig_3d_3d. A preview of the teachers' questions can be seen here: http://www.surveymonkey.com/s.aspx?sm=bqznzfw5IFRXRlft_2fH0SxA_3d_3d. Printed copies of the on-line surveys are also attached. All responses will remain anonymous.

Benefits for participation in this survey include opportunities to share information regarding principal evaluation practices. Your assistance in the survey will add to the limited research in this area. The aggregated results of this study will be used to inform principal training organizations. Aggregated results will also be shared with the Minnesota Department of Education, Minnesota Association of Secondary School Principals, and principal training colleges and universities. Furthermore, at the conclusion of the study, all participants will have the opportunity to view the survey results on-line.

If you have any questions or concerns regarding this study and would like to talk to someone other than the researcher, you are encouraged to contact the Research Subjects' Advocate Line, D528 Mayo, 420 Delaware St. Southeast, Minneapolis, Minnesota 55455; 612-625-1650. You can contact me at [address and phone number inserted].

Thank you for your time and assistance.

Sincerely,

Shirley Gregoire

University of Minnesota, Doctoral Candidate
Principal, St. Anthony Middle School in St. Anthony, MN

Informed Consent for Survey Participants

Dear Survey Respondent,

As part of the research for my doctoral degree in Educational Policy and Administration at the University of Minnesota, I am conducting a survey on the effectiveness of principals as evaluators of teachers for purposes of awarding merit pay. The study focuses specifically on those middle schools in Minnesota in which the principal's evaluation affects the awarding of merit pay. Your school is one of six schools that fit that parameter.

I would sincerely appreciate if you could take between 10 – 20 minutes to complete this online survey. All responses will remain anonymous and confidential. Data are being collected statewide and no school or individual principal or teacher is being tracked. Benefits for participation in this survey include opportunities to share information regarding principal evaluation practices. The results of this study will be used to inform principal training both in-service and pre-service. Results will also be shared with the Minnesota Department of Education, Minnesota Association of Secondary School Principals, and principal training organizations. Furthermore, at the conclusion of the study, all participants will have the opportunity to view the survey results on-line.

No risks are anticipated in this study outside those associated with the normal professional work day. Your superintendent has given permission for the teachers and principal at your school to participate in the study and all tenured teachers at participating schools are being invited to complete the survey. By clicking on the following link and completing this survey, you give your consent for the data to be used as part of the study.

Thank you for your time and assistance.

Sincerely,

Shirley Gregoire
University of Minnesota, Doctoral Candidate
Principal, St. Anthony Middle School in St. Anthony, MN

APPENDIX D

STATISTICAL TABLE

Two-sided *t*-test to Determine Significance in Importance

Average Score:	2.888889	2.907407	2.888889	2.833333	2.518519	2.759259	2.740741	2.555556	2.833333	2.87037	2.666667	2.648148
Statement #	1	2	3	4	5	6	7	8	9	10	11	12
1	1	0.7661	1	0.4437	**0.001041**	0.0896	0.0586	**0.0001**	0.4964	0.7844	**0.009195**	**0.00558**
2	0.7661	1	0.7092	0.2893	**0.002959**	**0.031**	**0.0055**	**7.25E-06**	0.2089	0.4846	**0.0011**	**0.0021**
3	1	0.7092	1	0.3706	**0.0002**	**0.0335**	**0.0444**	**0.0001**	0.4105	0.7661	**0.0092**	**0.0036**
4	0.4437	0.2893	0.3706	1	**8.48E-05**	0.2519	0.2793	**0.0043**	1	0.5686	**0.04854**	**0.0239**
5	**0.001041**	**0.002959**	**0.0002**	**8.48E-05**	1	**0.0036**	**0.038**	0.7489	**0.0011**	**0.0001**	0.1318	0.1803
6	0.0896	**0.031**	**0.0335**	0.2519	**0.0036**	1	0.7844	**0.01**	0.2519	0.083	0.1995	0.1822
7	0.0586	**0.0055**	**0.0444**	0.2793	**0.038**	0.7844	1	**0.001**	0.1676	**0.033**	0.3506	0.2551
8	**0.0001**	**7.25E-06**	**0.0001**	**0.0043**	0.7489	**0.01**	**0.001**	1	**0.0012**	**0.0002**	0.1592	0.2551
9	0.4964	0.2089	0.4105	1	**0.0011**	0.2519	0.1676	**0.0012**	1	0.4864	**0.0112**	**0.011**
10	0.7844	0.4846	0.7661	0.5686	**0.0001**	0.083	**0.033**	**0.0002**	0.4864	1	**0.0036**	**0.002**
11	**0.009195**	**0.0011**	**0.0092**	**0.04854**	0.1318	0.1995	0.3506	0.1592	**0.0112**	**0.0036**	1	0.811
12	**0.00558**	**0.0021**	**0.0036**	**0.0239**	0.1803	0.1822	0.2551	0.2551	**0.011**	**0.002**	0.811	1
13	0.1995	0.1095	0.2286	0.6417	**0.0123**	0.6417	0.472	**0.0056**	0.6739	0.2519	0.07	0.0733
14	0.2089	0.058	0.2089	0.799	**0.0082**	0.4105	0.2519	**0.0002**	0.799	0.3706	**0.0444**	**0.03771**
15	**0.0002**	**5.17E-06**	**0.0001**	**0.00066**	1	**0.008**	**0.002**	0.6417	**0.0004**	**5.29E-05**	0.088	0.1635
16	**5.27E-09**	**3.18E-10**	**8.24E-10**	**4.6E-09**	**0.0037**	**1.27E-07**	**4.55E-08**	**6.15E-05**	**1.18E-08**	**2.07E-09**	**4.65E-06**	**9E-06**
17	**6.37E-12**	**1.83E-11**	**1.87E-11**	**1.64E-11**	**2.48E-05**	**5.13E-09**	**1.17E-08**	**1.22E-05**	**8.22E-10**	**1.72E-12**	**6.06E-08**	**4.91E-07**
18	**0.001408**	**0.00057**	**0.00096**	**0.003**	0.7423	0.062	**0.04**	1	**0.0099**	**0.0004**	0.3079	0.3015
19	**0.011**	**0.0103**	**0.01696**	0.109	0.096	0.4964	0.6216	0.0586	0.109	**0.0111**	0.6417	0.472
20	**8.73E-06**	**4.39E-06**	**3.94E-06**	**5.17E-06**	0.5321	**0.0016**	**0.005**	0.2774	**0.0003**	**3.29E-05**	**0.027**	**0.0399**
21	**0.0149**	**0.0037**	**0.0149**	0.0882	0.1405	0.3988	0.4105	0.09	0.073	**0.006**	0.8296	0.6216
22	0.2893	0.133	0.1592	0.7844	**0.005**	0.4105	0.2519	**0.001**	0.799	0.4105	0.088	0.3077
23	**1.31E-07**	**9.74E-09**	**2.3E-08**	**1.17E-06**	0.063	**1.62E-05**	**4.39E-06**	**0.003**	**5.72E-07**	**5.27E-08**	**0.0003**	**0.0007**
24	**0.002959**	**3.26E-05**	**6.15E-05**	**0.0014**	0.8712	**0.005**	**0.0056**	0.5366	**0.0002**	**5.89E-05**	0.08321	0.0882
Significance #*	13	15	15	11	2	8	9	3	11	14	4	4

Average Score:	2.796296	2.814815	2.518519	2.166667	2.018519	2.555556	2.703704	2.444444	2.685185	2.814815	2.277778	2.5
Statement #*	13	14	15	16	17	18	19	20	21	22	23	24
1	0.1995	0.2089	**0.0002**	**5.27E-09**	**6.37E-12**	**0.001408**	**0.011**	**8.73E-06**	**0.0149**	0.2893	**1.31E-07**	**0.002959**
2	0.1095	0.058	**5.17E-06**	**3.18E-10**	**1.83E-11**	**0.00057**	**0.0103**	**4.39E-06**	**0.0037**	0.133	**9.74E-09**	**3.26E-05**
3	0.2286	0.2089	**0.0001**	**8.24E-10**	**1.87E-11**	**0.00096**	**0.01696**	**3.94E-06**	**0.0149**	0.1592	**2.3E-08**	**6.15E-05**
4	0.6417	0.799	**0.00066**	**4.6E-09**	**1.64E-11**	**0.003**	0.109	**5.17E-06**	0.0882	0.7844	**1.17E-06**	**0.0014**
5	**0.0123**	**0.0082**	1	**0.0037**	**2.48E-05**	0.7423	0.096	0.5321	0.1405	**0.005**	0.063	0.8712
6	0.6417	0.4105	**0.008**	**1.27E-07**	**5.13E-09**	0.062	0.4964	**0.0016**	0.3988	0.4105	**1.62E-05**	**0.005**
7	0.472	0.2519	**0.002**	**4.55E-08**	**1.17E-08**	**0.04**	0.6216	**0.005**	0.4105	0.2519	**4.39E-06**	**0.0056**
8	**0.0056**	**0.0002**	0.6417	**6.15E-05**	**1.22E-05**	1	0.0586	0.2774	0.09	**0.001**	**0.003**	0.5366
9	0.6739	0.799	**0.0004**	**1.18E-08**	**8.22E-10**	**0.0099**	0.109	**0.0003**	0.073	0.799	**5.72E-07**	**0.0002**
10	0.2519	0.3706	**5.29E-05**	**2.07E-09**	**1.72E-12**	**0.0004**	**0.0111**	**3.29E-05**	**0.006**	0.4105	**5.27E-08**	**5.89E-05**
11	0.07	**0.0444**	0.088	**4.65E-06**	**6.06E-08**	0.3079	0.6417	**0.027**	0.8296	0.088	**0.0003**	0.08321
12	0.0733	**0.03771**	0.1635	**9E-06**	**4.91E-07**	0.3015	0.472	**0.0399**	0.6216	0.3077	**0.0007**	0.0882
13	1	0.7423	**0.002**	**1.3E-07**	**1.3E-10**	**0.011**	0.2	**0.0004**	0.08321	0.799	**8.09E-06**	**0.005**
14	0.7423	1	**0.0004**	**2.7E-08**	**8.74E-10**	**0.005**	0.057	**2.71E-05**	0.0703	1	**8.47E-08**	**0.0004**
15	**0.002**	**0.0004**	1	**0.0004**	**8.52E-05**	0.7092	**0.049**	0.4193	**0.0377**	**0.0007**	**0.0412**	0.8547
16	**1.3E-07**	**2.7E-08**	**0.0004**	1	0.1723	**0.0009**	**2.32E-06**	**0.0078**	**1.14E-06**	**1.28E-09**	0.2774	**0.0014**
17	**1.3E-10**	**8.74E-10**	**8.52E-05**	0.1723	1	**7.4E-06**	**2.71E-08**	**0.0013**	**2.75E-08**	**3.63E-10**	**0.042**	**0.00014**
18	**0.011**	**0.005**	0.7092	**0.0009**	**7.4E-06**	1	0.088	0.2774	0.1278	**0.0021**	**0.0059**	0.5686
19	0.2	0.057	**0.049**	**2.32E-06**	**2.71E-08**	0.088	1	**0.0069**	0.811	0.1095	**7.96E-06**	**0.015**
20	**0.0004**	**2.71E-05**	0.4193	**0.0078**	**0.0013**	0.2774	**0.0069**	1	**0.027**	**5.89E-05**	0.1294	0.6166
21	0.08321	0.0703	**0.0377**	**1.14E-06**	**2.75E-08**	0.1278	0.811	**0.027**	1	0.09	**0.0004**	0.058
22	0.799	1	**0.0007**	**1.28E-09**	**3.63E-10**	**0.0021**	0.1095	**5.89E-05**	0.09	1	**2.68E-08**	**0.0007**
23	**8.09E-06**	**8.47E-08**	**0.0412**	0.2774	**0.042**	**0.0059**	**7.96E-06**	0.1294	**0.0004**	**2.68E-08**	1	**0.0092**
24	**0.005**	**0.0004**	0.8547	**0.0014**	**0.00014**	0.5686	**0.015**	0.6166	0.058	**0.0007**	**0.0092**	1
Significance #*	9	11	3	0	0	3	6	2	5	9	1	3

APPENDIX E

TABLE OF ATTRIBUTES, STANDARDS, AND GUIDELINE

Descriptor of Attributes, Standards, and Guideline Statement Text		
Survey Statement #	Standard Code	Statement Text
Propriety Attribute *Insures that evaluations are conducted legally, ethically, and with proper concern for the welfare of those involved in the evaluation.*		
	Propriety Standard #4 (P4)- *Interactions with Employees*	
1	P4	Conducted the evaluation in a respectful manner.
2	P4	Conducted the evaluation in a manner that I considered fair.
3	P4	Communicated the results clearly and objectively.
4	P4	Developed a relationship of mutual trust and understanding with me prior to the evaluation.
5	P4	Demonstrated a genuine interest in me as a person.
6	P4	Took into account my personal and professional needs.
	Propriety Standard #5 (P5)- *Comprehensive Evaluation*	
7	P5	Ensured I knew what would be assessed.
8	P5	Ensured I knew how the evaluation data would be collected.
9	P5	Described and justified the basis for interpretation of both positive and negative assessment information and results.
10	P5	Reported fully both strengths and weaknesses with supporting evidence.
11	P5	Supported continued, timely professional growth.
Utility Attribute *Insures that evaluations are conducted in a manner that insures evaluations are timely, informative, and useful.*		
	Utility Standard #4 (U4)- *Explicit Criteria*	
12	U4	Addressed only identified professional roles and responsibilities in the evaluation report and ensured that extraneous comments beyond the criteria were neither included nor accepted.
13	U4	Provided copies of written evaluation reports to me.
14	U4	Used the agreed upon criteria, providing a rationale and justification of evaluation findings.
Feasibility Attribute *Insures that evaluations are easily implemented, efficient in terms of time and resources, and are adequately funded.*		
	Feasibility Standard #1 (F1)- *Practical Procedures*	
15	F1	Delineated the procedures by which I could exercise my rights to review data about my performance.
16	F1	Encouraged me to suggest ways by which evaluation procedures could be made more efficient and useful.
Accuracy Attribute *Insures that evaluations are technically adequate so that the information generated can be used to make sound judgments.*		
	Accuracy Standard #1 (A1)- *Valid Judgments*	
17	A1	Considered my background and cultural experiences when interpreting performance.
18	A1	Ensured that my scoring was not influenced by factors irrelevant to my performance being evaluated (general impression or previous rating influences the present rating).
19	A1	Ensured that my summary conclusions which were derived from a series of assessments corresponded with the documented results.
	Accuracy Standard #8 (A8)- *Bias Identification and Management*	
20	A8	Obtained data and judgments from multiple sources to ensure validity and consistent indications of my performance.
21	A8	Allowed me to review data and participate in interpreting it where appropriate.
22	A8	Based my evaluations on defensible information with conclusions that were justifiable.
	Accuracy Standard #10 (A10)- *Justified Conclusions*	
23	A10	Generated, assessed, and reported plausible alternative explanations of findings and, if appropriate, indicated why these explanations should be discounted.
24	A10	Limited conclusions to those situations, time periods, contexts, and purposes for which the evaluation findings were applicable.