

Investigating a Teacher Evaluation System:
School Administrator and Teacher Perceptions of the
System's Standards of Effectiveness

by

Noelle A. Paufler

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved November 2014 by the
Graduate Supervisory Committee:

Audrey Amrein-Beardsley, Chair
David C. Berliner
Gustavo E. Fischman

ARIZONA STATE UNIVERSITY

December 2014

ABSTRACT

Increasing public criticism of traditional teacher evaluation systems based largely on classroom observations has spurred an unprecedented shift in the debate surrounding educational accountability policies, specifically about the purposes for and measures used to evaluate teachers. In response to growing public demand and associated federal mandates, states have been prompted to design and implement teacher evaluation systems that use increasingly available, statistically complex models (i.e., value-added) intended to isolate and measure the effects of individual teachers on student academic growth over time. The purpose of this study was to examine the perceptions of school administrators and teachers within one of the largest school districts in the state of Arizona with regards to the design and implementation of a federally-supported, state policy-directed teacher evaluation system based on professional practice and value-added measures. While much research has been conducted on teacher evaluation, few studies have examined teacher evaluation systems in context to better understand the standards of effectiveness used by school administrators and teachers to measure system effectiveness. The perceptions of school administrators and teachers, considering their lived experiences as the subjects of the nation's new and improved teacher evaluation systems in context, must be better understood if state and federal policymakers are to also better recognize and understand the consequences (intended and unintended) associated with the design and implementation of these systems in practice.

DEDICATION

This dissertation is dedicated to my husband, Tibor, in appreciation for his love and support throughout this journey. Although a few may have wondered silently why I would want to return to school *again*, he never questioned my interest or dampened my enthusiasm. Rather, he faithfully assumed extra household duties so that I could read another chapter or finish the next paper. For all the sacrifices, large and small, I am truly grateful. This mission could not have been accomplished without him, and I eagerly look forward to the next chapter in our lives.

ACKNOWLEDGMENTS

I wish to thank all of those who offered their support and encouragement to make this dissertation possible. First, to the school administrators and teachers who participated in the study, I sincerely appreciate your time, honesty, and trust. Without your willingness to candidly share your experiences and concerns, I could not have hoped to explore this topic and examine pertinent questions in a meaningful way. I am also grateful to district administration for the generous support and encouragement. Your assistance was instrumental to the success of this study.

I would also like to thank my dissertation committee: Dr. David Berliner, Dr. Gustavo Fischman, and Dr. Audrey Amrein-Beardsley for your time and support. I am grateful to each of you for the opportunity to work under your guidance and mentorship. Your willingness to share your wisdom and give honest feedback has been invaluable. Audrey, I could never have imagined the profound impact this experience would have on me as an educator and young scholar. The countless hours you spent answering questions, reviewing work, and offering professional advice made all the difference. You are a wonderful mentor and role model, and I will strive to pay it forward.

I am eternally grateful to my family for their unfailing love and support. First and foremost to my parents: words cannot express how deeply thankful I am for all your sacrifices over the years. You encouraged me to pursue my dreams, always confident that I could without ever insisting that I should. I am so proud to be your daughter. To Holly and Collin: thank you both for your positive outlook and willingness to listen to my daily updates even though I sometimes repeated myself. And to Emmett: even though you

cannot tell me in person, I know you share in my excitement. You were always one of my loudest cheerleaders.

Finally, I would like to thank my own teachers who year after year encouraged me to work hard and always do my best. Your love of learning and dedication to your students inspired me to become a teacher in the first place.

TABLE OF CONTENTS

	Page
LIST OF TABLES	xi
LIST OF FIGURES	xiii
CHAPTER	
1 INTRODUCTION	1
Background	1
Local Context	4
Purpose of the Study	8
School Reform as a Policy Cycle	9
Research Questions	11
Significance of the Study	11
Dissertation Overview	12
2 LITERATURE REVIEW	15
Teacher Supervision and Evaluation in Early America	15
Conflicting Views on Education	16
Evolution of Clinical Supervision	18
Clinical Supervision Model	18
Alternative Models of Supervision	19
Danielson Framework for Teaching	20
Role of the School Administrator	24
Influence on Teacher Quality	24
Effect of Classroom Observations	26

CHAPTER	Page
Evaluating Teachers Based on Student Achievement	27
History of Value-Added Modeling	28
Methodological and Pragmatic Issues in Teacher Evaluation	29
Value-Added Models	29
Validity	30
Reliability	34
Fairness	35
Clinical Supervision	37
Validity	38
Reliability	42
Fairness	45
Designing, Implementing, and Improving Teacher Evaluation	48
Using Teacher Evaluation Systems for High-Stakes Decisions	49
3 CONCEPTUAL FRAMEWORK	52
School Reform as Embedded in Contexts	52
School Reform as a Policy Cycle	53
Measuring the Effectiveness of School Reform	56
Symbolic Adaptation of School Reform	58
Understanding Stakeholder Perceptions	60
4 METHODS	62
Role of the Researcher	62
Pragmatic Paradigm Stance	62

CHAPTER	Page
Mixed Methods Research Design.....	63
Phase 1: School Administrator and Teacher Interviews	67
Participant Sampling	67
Interview Protocol	69
Data Collection	72
Data Management.....	73
Data Analysis.....	73
Phase 2: School Administrator and Teacher Surveys	75
Conducting a Census	75
Survey Instruments	76
Data Collection	81
Response Rates.....	82
Sample Representativeness.....	83
Data Management.....	84
Data Analysis.....	84
Quantitative Data	84
Qualitative Data	85
Validity	88
Researcher Lens.....	88
Participant Lens	90
Peer Reviewer Lens	95
Reliability	96

CHAPTER	Page
Generalizability	97
Study Limitations	98
5 RESULTS.....	100
Demographic Description of Survey Respondents	101
School Administrators	101
Teachers	103
Research Question 1: Purpose of Teacher Evaluation	107
Research Question 2: Intended Implementation	113
Transparency.....	113
Understanding System Components	115
Understanding Evaluation Processes	119
Research Question 2: Actual Implementation.....	125
Fidelity of Evaluation Processes	125
Evaluator Training and Objectivity.....	128
Time Spent in the Classroom	130
Research Question 3: Measuring System Effectiveness.....	131
Validity	131
Content-related Validity.....	131
Criterion-related Validity	137
Construct-related Validity.....	139
Consequential Validity.....	141
Reliability.....	142

CHAPTER	Page
Fairness	148
Intended and Unintended Consequences	153
Impact on Professional Practice.....	153
Impact on Student Achievement.....	157
Impact on Teacher Hiring and Retention	158
Impact on Public Perceptions	162
Summary of Results.....	164
6 FINDINGS AND CONCLUSIONS	165
Study Summary.....	165
Findings and Implications.....	167
Purpose.....	168
Assertion 1.....	168
Fidelity of Implementation.....	169
Assertion 2.....	169
Popularity.....	172
Assertion 3.....	173
Adaptiveness.....	174
Assertion 4.....	174
Longevity	176
Assertion 5.....	178
Conclusions	180
Recommendations for Further Research	183

CHAPTER	Page
REFERENCES.....	185
APPENDIX	
A ASSOCIATION FOR SUPERVISION AND CURRICULUM DEVELOPMENT: COPYRIGHT PERMISSION FOR USE	199
B ASSOCIATION FOR SUPERVISION AND CURRICULUM DEVELOPMENT: COPYRIGHT PERMISSION TO REPRINT	202
C MID-SOUTH EDUCATIONAL RESEARCH ASSOCIATION: COPYRIGHT PERMISSION TO REPRINT	205
D ARIZONA STATE UNIVERSITY INSTITUTIONAL REVIEW BOARD APPROVAL	207
E DISTRICT RESEARCH APPROVAL	210
F INTERVIEW PARTICIPATION LETTER.....	213
G PARTICIPANT INFORMED CONSENT FORM	216
H INTERVIEW REMINDER LETTER.....	219
I SCHOOL ADMINISTRATOR INTERVIEW PROTOCOL.....	221
J TEACHER INTERVIEW PROTOCOL	224
K SCHOOL ADMINISTRATOR SURVEY PROTOCOL.....	227
L TEACHER SURVEY PROTOCOL.....	236
M SCHOOL ADMINISTRATOR SURVEY PARTICIPATION LETTER	246
N TEACHER SURVEY PARTICIPATION LETTER.....	248
O SURVEY PARTICIPATION REMINDER LETTER	250
P CODE SHEETS.....	252

LIST OF TABLES

Table	Page
1. Qualitative Interview Analysis: Frequency of Themes by Position.....	75
2. Employment and Demographic Characteristics of School Administrators .	102
3. Employment Characteristics of Teachers.....	104
4. Demographic Characteristics of Teachers.....	106
5. Primary Reasons for Evaluating Teachers	112
6. Understanding of Teacher Evaluation System Component Calculations	119
7. Adequacy of District Communication.....	121
8. Helpfulness of Online Resources, Professional Development, and Communication with Others	123
9. 2012-2013 District Teacher Effectiveness Classification Report.....	124
10. Administrators Reported the Proportions of Teachers Who Completed or Participated in Evaluation Activities in the 2013-2014 School Year.....	125
11. Teachers Who Reported Completing or Participating in Evaluation Activities in the 2013-2014 School Year	126
12. Utility of Evaluation Activities for Improving Teacher Professional Practice	127
13. Evaluator Training, Objectivity, and Time Spent in the Classroom	128
14. Danielson Framework for Teaching.....	133
15. Considering Whether the District Should Add Non-test Information to the Teacher Evaluation System	136

Table	Page
16. Considering Whether the District Should Add Alternative Student Achievement or Learning Measures to the Teacher Evaluation System	137
17. Best Indicator of Effective Teaching.....	140
18. Weights Assigned by School Administrators and Teachers to Measures of Teacher Effectiveness.....	141
19. Teacher Overall Effectiveness Classification Labels	142
20. Improving the Teacher Evaluation System Using Multiple, Content-specific, External, and/or Peer Evaluators	145
21. Improving the Teacher Evaluation System through Additional Training and/or Danielson Framework for Teaching (FFT) Rubric Clarification.....	147
22. Adequacy and Fairness of the Teacher Evaluation System.....	150
23. Ability to Control and Improve Future Teacher Evaluation System Outcomes.....	153
24. Impact of the Teacher Evaluation System on Professional Practice	156
25. Administrators Reported the Impacts of the Teacher Evaluation System on Their Professional Practice	156
26. Impacts of the Teacher Evaluation System on Teacher Professional Practice	157
27. Impact of the Teacher Evaluation System on Student Academic Achievement and Learning.....	158

LIST OF FIGURES

Figure		Page
1.	A Blueprint for Teacher Evaluation	22
2.	Sequential Mixed Methods Research Design	65
3.	Alignment of Interview Protocols and Standards of Effectiveness.....	71
4.	Alignment of Survey Items and Standards of Effectiveness	79

CHAPTER 1

Introduction

The improvement of instructional quality has long been recognized in the educational community as the primary mechanism for increasing student learning. The traditional role of instructional supervision as a means of ensuring teacher quality has served as the foundation for teacher evaluation for more than two centuries in the United States (Cogan, 1973; Danielson, 2007; Goldhammer, 1969; Marzano, Frontier, Livingston, 2011; Tracy, 1995). Despite this sustained focus on hiring, developing, and retaining competent teachers in the classroom in order to promote student learning (Danielson & McGreal, 2000; Taylor & Tyler, 2012), teacher evaluation systems based almost entirely on supervision or classroom observations have been challenged in recent decades as inadequate measures of teacher effectiveness (Harris, 2011; Tucker & Stronge, 2005). As part of an unprecedented and fundamental shift in the discourse on accountability, the purpose for and measures used to evaluate teachers are at the forefront of education policy debates in states and school districts across the nation.

Background

Despite persistent efforts over the previous two centuries to improve the quality of schools in the United States by evaluating and developing the skills of teachers, the publication of *A Nation at Risk* in 1983 sounded the alarm among many Americans, insisting that the nation was purportedly “at risk” of imminent economic decline due to poor academic achievement. Among several recommendations outlined in the report, The National Commission on Excellence in Education (1983) proposed higher expectations of

professional competence for teachers in conjunction with increased salaries and more comprehensive evaluations:

Persons preparing to teach should be required to meet high educational standards, to demonstrate an aptitude for teaching, and to demonstrate competence in an academic discipline.... Salaries for the teaching profession should be increased and should be professionally competitive, market-sensitive, and performance-based. Salary, promotion, tenure, and retention decisions should be tied to an effective evaluation system that includes peer review so that superior teaching can be rewarded, average ones either improved or terminated. (p. 30)

In *Action for Excellence*, the Task Force on Education for Economic Growth, Education Commission of the States (ECS) (1983) renewed and reinforced the National Commission's emphasis on the importance of teacher competency, recommending the development and implementation of "systems for fairly and objectively measuring the effectiveness of teachers and rewarding outstanding performance" (p. 39). While the Task Force also acknowledged the need to professionalize teaching, its recommendations for evaluation reaffirmed the view that educational improvement hinges on "better teachers and better teaching" (Wise, Darling-Hammond, McLaughlin, & Bernstein, 1985, p. 62).

In response to the call for comprehensive teacher evaluation systems, Wise et al. (1985) examined supervision and teacher evaluation systems in 32 districts nationwide and conducted in-depth case studies in four of those districts. The models in place in the districts studied were predominantly developed by committees that included teachers, district and school administrators (e.g., principals), and union representatives. Four

primary problems were identified across the supervision and evaluation systems in use: 1) a lack of “sufficient resolve and competence” among principals to conduct evaluations, 2) teacher resistance to the feedback provided, 3) a lack of uniform evaluation processes, and 4) a lack of training for evaluators (Wise et al., 1985, p. 75). In addition, the researchers concluded that narrative evidence of teacher effectiveness was seen as less scientific, even by the teachers who preferred a more standardized approach. The researchers developed a series of recommendations based on five conclusions: 1) the evaluation system should suit the district’s goals, management style, conception of teaching, and community values; 2) administrative commitment to and resources for evaluation must supersede checklists and procedures; 3) the process must match the purpose of teacher evaluation in the district; 4) the utility of the system depends upon the efficient use of resources to achieve reliable, valid, and cost-effective evaluations; and 5) teacher responsibility for and participation in the process improves the quality of evaluation (Wise et al., 1985, p. 103-110). The researchers’ conclusions and associated recommendations were intended for use by districts to develop and implement successful teacher evaluation systems that were tailored to local needs.

Despite the strongly misgivings of critics Berliner and Biddle (1995) who compiled compelling evidence that claims of a crisis were intended to mislead and distort evidence of the accomplishments of public schools, these reports citing purportedly low academic performance in the United States as a precursor to economic decline spurred demands for national standards that would define the content taught to students in all public schools. Soon thereafter, and championed by President George H.W. Bush and three successors, the goals for standards-based educational reform were first outlined in

Goals 2000 and later incorporated into the updated Elementary and Secondary Education Act (ESEA) in 2002, under the moniker No Child Left Behind (NCLB). Mandating that all students demonstrate proficiency on state-determined standards in reading and mathematics by 2014, NCLB necessitated the development of large-scale standardized tests in every state for the purposes of measuring student learning and ultimately school quality (David & Cuban, 2010).

Since the passage of NCLB and provision of Race to the Top (RttT) and Teacher Incentive Fund (TIP) grants by the United States Department of Education (2009, 2010), states have been prompted to develop and implement accountability systems to measure teacher, principal, and school effectiveness relying at least in part on student performance on state-level tests (Amrein-Beardsley 2008, 2014; Braun, 2005; Corcoran, 2010). RttT required states to provide evidence of compliance in the development and implementation of such an accountability system in order to be eligible for grant funds (United States Department of Education, 2009). As a result, states across the nation are developing and implementing such systems based on quantitative measures of teacher and school effectiveness with high-stakes consequences (Amrein-Beardsley & Collins, 2012; Berliner, 2014; Corcoran, 2010; Weisberg, Sexton, Mulhern, & Keeling, 2009).

Local Context

Public school districts in the state of Arizona have similarly adopted teacher evaluation systems aligned to current federal policy. In 2010, the Arizona state legislature modified existing evaluation policies with the passage of Senate Bill 1040 (Arizona Revised Statutes §15-203 (A) (38)) which coincided with the state's application for RttT funds. In compliance with Senate Bill 1040, the Arizona State Board of Education (ADE)

provided a framework through which all public school districts and charter schools in the state beginning in the 2012-2013 school year would be required to annually evaluate teacher effectiveness according to four performance classifications (Highly Effective, Effective, Developing, and Ineffective). This was to be done using both measures of student academic progress and professional practice. Each district was required to design and implement an evaluation system in which multiple measurements of the academic progress of students in each teacher's classroom would comprise between 33 and 50 percent of his or her evaluation rating (Arizona Department of Education [ADE], 2011). Additionally, between 50 and 67 percent of each teacher's rating must be based on multiple measurements of instructional quality through classroom observations (ADE, 2011). As such, school districts and charter schools in Arizona began the process of designing and implementing newly aligned teacher evaluation systems.

This study was conducted in one such large public school district in Arizona. As a result of these policy changes, the district designed and began implementing a new teacher evaluation system. Through a collaborative effort involving teachers, principals, district administration, curriculum and instruction specialists, the district teachers union, research staff, and others (as members of the Teacher Evaluation Committee), the district developed a model in the 2012-2013 school year that comprises both measures of student academic progress and professional practice. The district's model is aligned to the state policy-directed framework for evaluation and reflective of the larger national policy trends in accountability systems. As stated in the district's Certified Evaluation Process Handbook, the model is intended "to enhance teaching and student achievement through targeted professional development and data-informed decision making" as well as to

“bring consistency, common understanding and reflective dialogue to teaching and learning” as outlined in the following district objectives:

- 1) Providing a common district wide definition of effective teaching
- 2) Embracing meaningful discussion and collaboration about teaching practices
- 3) Focusing on continuous growth for all teachers
- 4) Identifying and emphasizing strategies have the greatest impact on student learning.

All certified staff members in the district (who will be subsequently referred to as “teachers”) are evaluated using this model including: elementary and high school classroom teachers (general and special education), instructional support staff (e.g., instructional coaches, reading and mathematics interventionists), counselors, and related services staff (e.g., psychologists, speech pathologists, etc.).

As part of the evaluation model, each teacher receives a teacher evaluation score (on a scale of 1 to 100). This score is a composite of two weighted scores: a professional practice score (67.0%) as determined by the teacher’s performance on the Danielson Framework for Teaching (FFT; Danielson, 2013) and a student academic progress score (33.0%) that is calculated through a value-added model using data from approved student achievement tests (i.e., Arizona’s Instrument to Measure Standards [AIMS] tests in reading, mathematics, and science). Next, each teacher is given a performance group assignment (on a scale of 1 to 4 from lowest to highest) based on his or her teacher evaluation score. These performance group assignment scores were determined by the district’s Teacher Evaluation Committee through a formal standards setting process. Lastly, each teacher receives an overall effectiveness classification corresponding to his

or her performance group assignment (i.e., 4 = Highly Effective, 3 = Effective, and 2 or 1 = Developing or Ineffective based on certain criteria). Specifically, any teacher who receives one or more “Unsatisfactory” ratings on the Danielson FFT and/or any continuing teacher (i.e., with four or more years of experience) who receives four or more “Basic” ratings is classified as Ineffective. Teachers who meet the above criteria at any time during the school year or who are otherwise identified by school administration are provided specific supports via a formalized plan of improvement. All teachers can access their individual historical evaluation data (e.g., evidence and ratings on the Danielson FFT, value-added score, overall effectiveness classification) through an internal online system (referred to as the Comprehensive Evaluation System [CES]).

Teachers in grades 3-8 for whom achievement data are available for their individual students or their content area (e.g., elementary self-contained classroom teachers) are considered part of Group A and receive a value-added score based on their students’ scores on AIMS reading, mathematics, science, or a combination of these. Teachers for whom this is not the case (e.g., secondary teachers, elementary special area teachers [i.e., art, music, and physical education]) are considered part of Group B and receive a value-added score based on grade- or school-level rather than individual data. In certain limited situations, teachers are assigned to Group A/B if they teach multiple content areas such that scores are only available at the student level for one of those areas. For example, a teacher in grade 7 who is assigned to both mathematics and science would be considered a Group A/B teacher as the AIMS science test is only administered to students in grades 4, 8, and 10. For the purposes of this study, teachers in Group A/B will be considered part of Group B.

As part of the model development process, pilot data were gathered in the 2012-2013 school year and used for the purposes of educating teachers about the process, making decision rules for the performance group assignments, etc. The model was formally utilized to evaluate approximately 1,400 classroom teachers and other certified staff in the district in the 2013-2014 school year.

Purpose of the Study

The purpose of this study was to examine the perceptions of elementary school administrators (i.e., principals and assistant principals) and classroom teachers (i.e., Groups A and B) regarding the new teacher evaluation system in place. While much research has been conducted on teacher evaluation, few studies have directly examined variations in the perceptions of stakeholders in a local context with regards to the purpose(s) of and implementation processes for a new teacher evaluation system. In addition, the ways in which the various stakeholders define and measure the effectiveness of their teacher evaluation system has not been fully explored as situated within a larger state policy-directed evaluation framework.

I sought to better understand how these recipients of, and actors within, the evaluation system thus far understand, define, and measure its effectiveness and overall “value-added.” Specifically, I investigated the extent to which their district system is aligned to the state policy-directed teacher evaluation framework in terms of the following: its *purpose*, *fidelity* of implementation, *popularity* among actors within the system, *adaptiveness* as part of professional practice, and *longevity* in the policy cycle (Cuban, 1998). These five standards of effectiveness provided the conceptual framework

through which school administrator and teacher perceptions were examined (Cuban, 1998) and will be discussed in greater depth.

School Reform as a Policy Cycle

Much research has been conducted in recent decades concerning school reform in general, as well as the reformation of schools as a policy process. Tyack and Cuban (1995) described education reform as occurring in cycles, including phases of policy talk, action, and implementation. Although many within the education profession have observed and criticized seemingly repetitious calls for the same or very similar reforms, Tyack and Cuban (1995) argue that this policy cycle occurs in different contexts over time as steady, albeit slow changes in schools as educational institutions reframe surrounding conversations. But in order to determine whether a reform has been successful, one must ask how success is to be measured.

Cuban (1998; see also Tyack & Cuban, 1995) also argued that one must inquire as to 1) whether the goals of a program were achieved (effectiveness), 2) to what extent the program was popular, and 3) whether the program was implemented with fidelity (p. 456-458). While these standards of success typically rely on quantitative results (e.g., students' standardized test scores), the use of these three standards to measure effectiveness and the ways in which local players understand and perhaps actively (de)legitimize the reforms of focus can serve as a useful approach to studying education reform. In addition, teachers often seek to alter and adapt reforms during implementation (Cuban, 1998). This standard of adaptiveness is also considered essential in order for a reform to meet the other most important standard for practitioners—that of longevity

(Cuban, 1998). Overall, in order for a reform to be considered a success to most teachers, it must outlast the next cycle of policy talk.

In this regard, teacher evaluation systems may be intended to legitimize the teacher as a professional and act as a symbol of credibility for the institution as having met its social mandate; however, this may not necessarily preclude teachers or school administrators from participating in or adapting reform activities such as those that are ancillary in a ceremonial or symbolic way (Popkewitz, Tabachnick, & Wehlage, 1982). Symbols, slogans, and rituals with regards to reform describe the meaning of “potential actions” but may not necessarily describe what is actually happening in practice (Popkewitz et al., 1982, p. 20-21). Based on this argument, reform in general may serve to legitimize schooling as an institution, to some extent protecting the same institution from public scrutiny. As such, school reform may in reality “conserve rather than change” procedures, rules, and practices through symbolic actions that may not reflect real activities (Popkewitz et al., 1982, p. 21).

The use of the policy cycle as a framework to better understand teacher evaluation at the district level offered a means of situating personal professional processes and understandings within their appropriate institutional structures. The aforementioned language used by Cuban (1998) to describe the various standards of effectiveness (i.e., purpose, fidelity of implementation, popularity, adaptiveness, and longevity) served as the platform for the research questions in this study. These questions examined the standards as defined by the different actors most pertinent here (i.e., school administrators and teachers).

Research Questions

Using the aforementioned conceptual framework, I generated the following overarching research questions relating to the perceptions of elementary school administrators and classroom teachers:

- 1) What do stakeholders perceive as the purpose and goals of the locally-developed teacher evaluation system in use in their district?
- 2) How do stakeholders describe the intended and actual implementation processes for the teacher evaluation system?
- 3) How do stakeholders measure the effectiveness of the teacher evaluation system based on their understandings of the purpose/goals as well as the intended and actual implementation processes?
- 4) To what extent do perceptions of the purpose/goals, descriptions of implementation, and measures of effectiveness vary across stakeholder groups?

Significance of the Study

A better understanding of variations in these perceptions and implications for continued use of the system in that context has been of use to other stakeholders who continuously seek to facilitate dialogue between and among groups (e.g., district administrators, school leadership, teachers, etc.) with regards to the purpose(s) of the system, ways to implement the system most efficiently, and means by which to measure system effectiveness. Dialogue in these areas has and ideally will continue to provide stakeholders with opportunities to more closely examine the disparaging impact of state and federal evaluation policies on various groups, critique the design and implementation

of the evaluation system in the context of their district and school(s), and adapt the system to the extent possible within the state policy-directed framework provided. Thus far, analyses of perceptual variations among and within groups have informed district leadership in their efforts to improve system implementation processes and, ideally, will contribute to the literature on accountability and evaluation as state and federal policy changes necessitate the design, implementation, and evaluation of systems in varied contexts across the state and nation in a relatively short period of time.

Further research on teacher evaluation is warranted, especially given the ongoing, contentious debate occurring throughout the nation among politicians and policymakers, educational researchers and other scholars (e.g., econometricians), journalists and other popular press “experts,” educators, and the general public (often informed by politicians, policymakers, and the media). The perceptions of school administrators and teachers, considering their lived experiences as the subjects of the nation’s “new and improved” teacher evaluation systems in context, must be better understood if state and federal policymakers are to better recognize and understand the consequences (intended and unintended) associated with the design and implementation of these systems in practice. This study will contribute to the growing body of evidence needed if researchers are to help to inform and ultimately make substantive policy changes that are themselves effective in encouraging the recruitment, retention, and promotion of the best teachers.

Dissertation Overview

In Chapter 2, I provide a historical overview of teacher evaluation in the United States, focusing on the progression of early clinical supervision to more contemporary models such as the Danielson FFT. In addition, I discuss the role of the school

administrator as teacher evaluator and the effect of classroom observations on teacher performance. I also present an historical overview of value-added modeling as well as current policy trends regarding its use for evaluating teachers. Finally, I examine methodological and pragmatic issues in the use of value-added and classroom observation models by reviewing related empirical research studies.

In Chapter 3, I explain the conceptual frameworks upon which the study research questions, design, and analytic approaches were based. Specifically, I discuss the phases of the policy cycle (i.e., talk, action, and implementation) and their applicability to policy trends in teacher accountability (Tyack & Cuban, 1995). I also examine the standards of effectiveness developed by Cuban (1998) (i.e., purpose, fidelity of implementation, popularity, adaptiveness, and longevity) and the concept of symbolic adaptation (Popkewitz et al., 1982) as potentially useful in understanding school administrator and teacher perceptions in the context of this study.

In Chapter 4, I describe the sequential mixed methods research design developed for this study. After discussing processes for instrument development as well as data collection, management, and analyses, I outline additional research activities completed to determine the validity, reliability, and generalizability of study results. I also address the limitations of the study.

In Chapter 5, I present the study results, integrating and organizing interview and survey data by research question. For each question, I discuss results thematically as appropriate with exemplary qualitative and descriptive quantitative evidence. In an effort to describe participants' experiences in an authentic, meaningful way, I rely upon their words to contextualize descriptive survey results.

In Chapter 6, I summarize the study before presenting and supporting assertions for each standard of effectiveness in the conceptual framework. I also discuss the applicability of the policy cycle and symbolic adaptation as concepts in the context of this study. In conclusion, I address the potential contribution of the study to inform district, state, and national policymakers with regards to the intended and unintended consequences of teacher accountability policies in practice and propose additional areas for teacher evaluation research.

CHAPTER 2

Literature Review

In this chapter, I illustrate the historical trajectory of teacher evaluation in the United States over the past two centuries. I provide an overview of the traditional role of clinical supervision in improving teacher quality, based in part on the work of Marzano et al. (2011; see Appendix A for copyright permission). I then introduce value-added modeling in the context of current policy trends in accountability. In addition, I review the most significant methodological and pragmatic issues associated with value-added and classroom observation models as situated within the literature.

Teacher Supervision and Evaluation in Early America

In colonial America during the 18th century, town governments and clergy provided local supervision of teachers, a responsibility often delegated to individuals or committees who had sole authority over hiring criteria and retention (Tracy, 1995, see also Marzano et al., 2011). As might be expected, feedback to teachers varied considerably. As part of the common schooling movement in the next century, more structured education systems were established in large urban areas (Marzano et al., 2011). Teachers with discipline-specific expertise and administrators with the ability to assume managerial responsibilities were sought to staff the schools (Marzano et al., 2011; Tyack & Cuban, 1995). Given the emerging view of teachers and administrators as professionals, clergy were no longer seen as qualified for teacher supervision (Tracy, 1995; see also Marzano et al., 2011). Over the next few decades, the importance of teachers' pedagogical skills in providing quality instruction, and subsequently, the need

for teacher supervision and more specific feedback was recognized, although not necessarily qualified (Tracy, 1995, p. 323; see also Marzano et al., 2011).

Conflicting Views on Education

Two dominant and often adversarial views on education emerged in the early 20th century: one based on the writings of Dewey whereby democracy served as the linchpin in human development and another more scientific conceptual understanding of education based on the work of Taylor whereby teaching was intended to prepare future workers (Marzano et al., 2011, p. 14). Dewey (1938, 1973) advocated for the utilization of schools as spaces to cultivate democratic values in students as citizens, suggesting that students would only be prepared to be active citizens if their schooling was student-centered, applicable to the real world, differentiated based on their needs, and interdisciplinary in nature (see also Marzano et al., 2011). In contrast, Taylor (1911/1998) influenced K-12 education practices by arguing that the measurement of factory workers' behaviors served as the primary mechanism for increasing production and insisting that the best method for completing tasks should be determined by level of efficiency (see also Marzano et al., 2011).

Thorndike, Cubberley, and others soon advanced measurement as a means by which to improve schooling. Cubberley (1929) expanded upon Taylor's concept of mechanized schooling, comparing public schools to factories in need of appropriate tools, specialized processes, and measures of efficiency:

Our schools are, in a sense, factories in which the raw products (children) are to be shaped and fashioned into products to meet the various demands of life. The specifications for manufacturing come from the demands of twentieth century

civilization, and it is the business of the school to build its pupils according to the specifications laid down. (p. 338)

Cubberley (1929) further argued that teachers should be provided with detailed feedback from their administrators when observed in the classroom as a means of increasing efficiency in instruction and output in terms of student performance (see also Marzano et al., 2011). In addition, Wetzel (1929) argued that teachers' strategies and behaviors should also be used to measure teacher and ultimately school quality through measures of students' aptitude, clear objectives and content standards, and reliable measures of student learning (see also Marzano et al., 2011). Throughout the first half of the 20th century, the debate continued between Dewey's ideas about the purpose of education and the demands of Cubberley and Wetzel that data be used to provide feedback and measure teacher, school, and district effectiveness (Marzano et al., 2011).

Almost immediately following the end of World War II, the dialogue about teachers shifted to emphasize their importance as individuals (Marzano et al., 2011). While the focus on teachers narrowed somewhat, Swearingen (1946) suggested that supervisors' responsibilities were expanded to include four areas related to teacher competency and evaluation: the curriculum, teaching personnel, the teaching/learning situation, and the emotional quality of the classroom. While the additional responsibilities of supervisors did not likely lead to increased efficiency, the importance placed on classroom observations as a means of providing feedback to teachers was invaluable (Marzano et al., 2011).

Evolution of Clinical Supervision

Clinical supervision model. The clinical supervision model has been one of the most rapidly adopted practices in the field. Beginning in the 1950s, Morris Cogan (professor and mentor in a teaching program at Harvard University) developed a model with the help of his students based on that used to supervise medical students completing their residency, emphasizing the importance of observation and discussion (Goldhammer, 1969). Goldhammer (1969) defined clinical supervision and explained its value, emphasizing the importance of the face-to-face relationships that must develop between supervisors and teachers in addition to the need for observations of actual professional behavior (p. 54). He further argued that clinical supervision is intended to incentivize and prepare teachers to engage in self-supervision and to supervise their colleagues. In addition, he suggested that the value of supervision increases as teachers become more skilled in their craft (p. 55). Goldhammer (1969; see also Marzano et al., 2011) outlined a cycle of supervision with five sequences (stages):

1. Preobservation Conference: The teacher and supervisor communicate and agree upon the purpose of and plan for the observation (p. 57-61).
2. Classroom Observation: The supervisor observes the teacher while engaged in professional behavior, namely to help the teacher “test reality” with regards to perceptions about his or her own practice. This is intended to increase the teacher’s independence, objectivity, and awareness and in turn prompt further self-reflection (p. 63).

3. Analysis: The data gathered during the observation must be synthesized for use by teachers to evaluate their own instruction. The strategy for the supervision conference to follow should be clearly outlined (p. 63).
4. Supervision Conference: This stage is intended to provide the teacher and supervisor with an opportunity to discuss his or her analysis of the behavior observed and, essentially, empower the teacher to self-reflect (p. 69).
5. Post-Conference Analysis: The supervisor also analyzes his or her own professional behavior, assessing the productivity of the supervision and identifying areas in need of change for future cycles (p. 71).

Cogan (1973), Goldhammer's student at Harvard, later expanded upon his work by identifying for supervisors the specific classroom behaviors or "critical incidents" that could be detrimental to student learning (p. 172). Citing supervision as important to improving teachers as professionals, Cogan (1973) also cautioned that the supervisor's personal teaching philosophy could inhibit him or her in dialogue with teachers about their practices (see also Marzano et al., 2011). Over time, the clinical supervision model has arguably changed (and deteriorated somewhat) from Goldhammer's original vision to a series of prescribed steps (Marzano et al., 2011). Goldhammer (1969) had not described any specific attributes of quality instruction, perhaps adding to the evolution (and ultimately confusion) of the model as a mechanism for evaluating teachers.

Alternative models of supervision. Over the next decade, alternative models of supervision emerged in response to the narrowly-defined uses of clinical supervision. Glatthorn (1984) argued that teachers should be empowered to choose from among four methods of evaluation based on their individual needs: 1) clinical supervision, 2)

collaboration with a colleague in a cooperative development program, 3) self-directed professional development, or 4) administrative monitoring. He suggested that clinical supervision would be most appropriate for beginning teachers and those experienced teachers who are struggling in the classroom.

McGreal (1983) also guided school districts seeking to examine their current evaluation system and develop alternatives. He emphasized the importance of a common understanding of the purpose of an evaluation system in the local context and the development of a system aligned to that purpose. Furthermore, he argued that the attitude of teachers and supervisors is critical to creating an effective system, and as a result, the process must facilitate collaboration between the two groups (p. 41). In the fourth edition of *Supervision of Instruction: A Developmental Approach* (originally published in 1985), Glickman, Gordon, and Ross-Gordon (1998) advocated for a differentiated approach through which the four supervisory behaviors (directive control, directive informational, collaborative, and nondirective) are appropriately matched with each teacher's developmental level, expertise, and commitment.

Danielson Framework for Teaching

In conjunction with larger changes in education policy over the past few decades, measures of teacher competence have shifted from teacher behavior to student achievement, and in turn, from clinical supervision to evaluation. Danielson published her seminal work, titled *Enhancing Professional Practice: A Framework for Teaching*, in 1996 (updated in 2007) based upon her experiences at the Educational Testing Service (ETS). The 2013 edition of the model is comprised of four domains of equal importance that are also aligned to the Interstate New Teacher Assessment and Support Consortium

(InTASC) standards (Council of Chief State School Officers, 2011): 1) Planning and Preparation, 2) the Classroom Environment, 3) Instruction, and 4) Professional Responsibilities. Each domain includes two to five components (22 in total) that are subdivided into specific, observable tasks or elements (76 in total) rather than statements about teachers' beliefs or values. Teachers are evaluated across all components according to four levels of proficiency (Unsatisfactory, Basic, Proficient, and Distinguished).

Danielson's model (see Figure 1) captures the multifaceted nature of teaching, provides a language for dialogue about teacher competence, and serves as a framework for teacher self-assessment and reflection (Marzano et al., 2011). With regards to improving instruction, Danielson's model arguably measures the construct of teacher quality to the extent that teacher quality is evidenced by observations of the specific behaviors included in each domain.

A Blueprint for Teacher Evaluation

Components of Professional Practice

<p><i>Domain 1: Planning and Preparation</i></p> <p>This domain includes comprehensive understanding of the content to be taught, knowledge of the student's backgrounds, and designing instruction and assessment.</p> <ul style="list-style-type: none">1a. Demonstrating knowledge of content and pedagogy1b. Demonstrating knowledge of students1c. Setting instructional outcomes1d. Demonstrating knowledge of resources1e. Designing coherent instruction1f. Designing student assessments	<p><i>Domain 2: The Classroom Environment</i></p> <p>This domain addresses the teacher's skill in establishing an environment conducive to learning, including both the physical and interpersonal aspects of the environment.</p> <ul style="list-style-type: none">2a. Creating an environment of respect and rapport2b. Establishing a culture for learning2c. Managing classroom procedures2d. Managing student behavior2e. Organizing physical space
<p><i>Domain 4: Professional Responsibilities</i></p> <p>This domain addresses a teacher's additional professional responsibilities, including self-assessment and reflection, communication with parents, participating in ongoing professional development, and contributing to the school and district environment.</p> <ul style="list-style-type: none">4a. Reflecting on teaching4b. Maintaining accurate records4c. Communicating with families4d. Participating in the professional community4e. Growing and developing professionally4f. Showing professionalism	<p><i>Domain 3: Instruction</i></p> <p>This domain is concerned with the teacher's skill in engaging students in learning the content, and includes the wide range of instructional strategies that enable students to learn.</p> <ul style="list-style-type: none">3a. Communicating with students3b. Using questioning and discussion techniques3c. Engaging students in learning3d. Using assessment in instruction3e. Demonstrating flexibility and responsiveness

Figure 1. Adapted from *Enhancing Professional Practice: A Framework for Teaching* by C. Danielson. Copyright 1996 by Association for Supervision and Curriculum Development. Reprinted with permission of the original copyright holder (see Appendix B).

It is important to note that Danielson (2007) focused on the role of supervision as a means of improving instruction rather than a system of evaluation. Danielson (2007) explained the purpose and design of the framework as follows:

The framework for teaching is based on important assumptions about what is important for students to learn, the nature of learning and how to promote it, the purposeful nature of teaching, and the nature of professionalism. The framework for teaching also has a number of important features: it is comprehensive, grounded in research, public, generic, coherent in structure, and independent of any particular teaching methodology. (p. 25)

Danielson (2007) further discussed the use of the framework for supervision and evaluation, noting the importance of a clear, research-based definition of teaching that reflects the “professional wisdom” of those who will implement the evaluation system (p. 177). These criteria should be made known to teachers in advance so that they have an opportunity to gather evidence related to each (Danielson, 2007). In addition, administrators must be adequately trained to “make consistent judgments based on evidence of practice...” (p. 177), and teachers must understand the criteria so that they can provide evidence of their skills (Danielson, 2007). It is important to consider, however, that teacher quality could arguably comprise other domains as well that are not included in the framework and that observations of behaviors currently outlined in each domain are only intended as a sampling that is used to generalize about the average or typical behaviors of any given teacher.

Danielson and McGreal (2000) also advised school districts seeking to build new evaluation systems. According to Danielson and McGreal (2000), districts should ensure

that their system is directly linked to the school/district mission, the development of the system is an ongoing process, the system emphasizes student outcomes, and there is a commitment to allocating the necessary resources for system success (p. 18-19).

Role of the School Administrator

Principals (and assistant principals for the purposes of this study) certainly play multiple roles in the complex organizational and instructional environments of their schools, arguably the most important of which is ensuring the high quality of instruction (Donaldson, 2011). The manner in which principals hire, assign, evaluate, and develop the professional capacities of teachers can insignificantly impact teacher and ultimately instructional quality (Donaldson, 2011). The increasing focus of state and federal policymakers, universities, foundations, and perhaps most importantly, school districts on teacher talent development through human capital initiatives is promising; however, further research is needed examining the role of the principal in this process, specifically in terms of raising teacher quality at his or her school (Donaldson, 2011).

Influence on teacher quality. The processes principals used to hire, assign, evaluate, and provide professional development opportunities to teachers can vary considerably by school site (Donaldson, 2011). Donaldson (2011) studied the processes used by 30 principals in public and charter schools in two northeastern states to determine what factors influenced principals in their above-referenced tasks, the constraints and opportunities affecting their completion of these tasks, and the differences in processes across various contexts. The researcher determined that some principals exercised more direct control over human capital functions (e.g., hiring, assigning, evaluating, or dismissing teachers) than others (Donaldson, 2011). In addition, principals reported a

variety of constraints in performing these tasks, ranging from economic influences to contractual limitations (Donaldson, 2011). Interestingly, the extent to which principals felt constrained in these areas varied little between public and charter schools. Principals in both cases who reported the fewest barriers in performing these tasks supervised schools that in general were smaller, served elementary students, exhibited strong local identities (according to their principals), and enjoyed widespread district-level support (Donaldson, 2011).

Donaldson (2011) concluded that principals in the sample “conceived of evaluation as serving two main purposes: first, to improve instruction and, second, to identify poorly performing teachers for intervention and, potentially, dismissal” (p. 17). Despite these seemingly clear objectives, principals in the study indicated that evaluation rarely achieved these outcomes (Donaldson, 2011). According to study findings, principals reported four primary constraints on their ability to effectively evaluate and recommend the dismissal of teachers who consistently performed below expectations: 1) lack of time, 2) limited opportunities to observe and document instruction representative of typical performance, 3) inadequate observation instruments, and 4) school culture (Donaldson, 2011). While most principals acknowledged the importance of formal observations, they also characterized the instruction that they witnessed as “staged” and frequently praised informal observations as more informative (Donaldson, 2011). Principals also noted that they already lacked sufficient time to conduct frequent informal observations let alone to provide substantive feedback. In addition, many principals felt that the observation instrument used in their district was inadequate, specifically that the instrument was cumbersome with regards to paperwork and often binary in nature (e.g.,

“meets standard” or “does not meet standard”) (Donaldson, 2011, p. 22-23). Lastly, principals reported that although they were empowered to observe and even dismiss tenured teachers, they hesitated to do so, most often citing school culture as discouraging the practice (Donaldson, 2011). In this study, principals reported a variety of barriers to raising teacher quality through evaluation and dismissal, although they acknowledged responsibility for these and other human capital functions in their schools (Donaldson, 2011).

Effect of classroom observations. The effect of principals’ (and assistant principals’) evaluations based on classroom observations and other similar teacher performance measures over time has been examined in the literature as well. The use of formal observations by principals to improve teacher quality may be less evident in the short term but certainly can affect change in mid-career performance when applied consistently alongside professional development training and other human capital investments (Taylor & Tyler, 2012). Taylor and Tyler (2012) examined the effect of teacher evaluation over time as measured by student achievement, specifically to determine whether evaluation improves teacher performance during the period in which the evaluation occurs, and also whether past evaluation improves teacher performance even after the teacher is no longer evaluated. Taylor and Tyler (2012) used Teacher Evaluation System (TES) data from Cincinnati Public Schools whereby teachers’ professional practices were measured through multiple classroom observations and a review of work products not related to student test scores. The researchers determined that “high-quality, classroom-observation-based evaluation improves mid-career teacher performance both during the period of evaluation and in subsequent years, though the

estimated improvements during evaluation are not always robust” (Taylor & Tyler, 2012, p. 3). Study findings suggested that formal classroom observations conducted by principals as part of teacher evaluation systems are an important tool for improving performance over time even if the effects are not immediately evident (Taylor & Tyler, 2012). These results have implications for school leadership as most teacher evaluation systems rely upon multiple measures of performance including principals’ formal evaluations based on classroom observation (Taylor & Tyler, 2012).

Evaluating Teachers Based on Student Achievement

In response to the increasing public demand for school and teacher accountability systems linked to student learning, statistically complex measures of accountability are now being used to isolate the educational output of individual teachers (Amrein-Beardsley, 2008, 2014; Harris, 2011; Papay, 2010). Value-added models (VAMs) are used to measure the effect of a teacher on his or her students’ learning from one year to the next using their scores on large-scale standardized tests (Braun, 2005; Scherrer, 2011). Unlike traditional snapshot measures of an individual student’s achievement at a single point in time or that of different cohorts of students at two points in time, however, VAM estimates are intended to measure student growth (Baker et al., 2010; Harris, 2009; Hershberg, Simon, & Lea-Kruger, 2004). In order to isolate the effect of the individual teacher from other factors that may impact a student’s growth, VAMs predict the student’s performance on a test using variables such as student background characteristics (e.g., racial or ethnic background, socioeconomic status, English language proficiency, special education needs, etc.) and prior achievement (Goe, 2008; Harris, 2011; Newton, Darling-Hammond, Haertel, & Thomas, 2010; Scherrer, 2011).

Attributing the difference between the predicted and actual performance of the student on the test as a measure of the “value-added” by his or her teacher (Goe, 2008; Scherrer, 2011), VAMs purport to identify (in)effective teachers and schools (Braun, 2005).

History of value-added modeling. Value-added modeling was first applied to education by Tennessee statistician Dr. William Sanders in the 1980s following its use in the field of agriculture genetics (Hong, 2010; Schaeffer, 2004). Convinced that a VAM could be used to improve teacher accountability, Sanders appealed directly to the governor of Tennessee for the rights to student test score data in Knox County Schools (Hong, 2010; Schaeffer, 2004). The state legislature soon thereafter, as based on Sanders’ preliminary evidence, adopted the VAM as the “methodology of choice” for measurements of district, school, teacher, and student performance (Hong, 2010, p. 3). Originally named the Tennessee Value-Added Assessment System (TVAAS), the model became a prototype for sweeping national reform in education accountability. Largely funded by the United States Department of Education, value-added pilot programs were developed in North Carolina, Arkansas, Delaware, and Florida by 2006 (Amrein-Beardsley, 2008). Five more states were expected to receive growth model project grants in the following year (Amrein-Beardsley, 2008). Bolstered by nearly \$100 million per year in federal funding for a four year period, participating states were expected to warehouse student test score data and incorporate value-added outputs into teacher evaluations (Amrein-Beardsley, 2008).

Although VAMs were already adopted in several states, they first drew widespread criticism after *The L.A. Times* published the results of a statistical analysis of student test data (Felch, Song, & Smith, 2010). Intending to provide information to the

public about elementary schools and teachers in the Los Angeles Unified School District (LAUSD), the newspaper chose to identify each teacher by name and disclose his or her value-added scores (Felch, Song, & Smith, 2010). This decision sparked a national debate regarding the strengths and limitations of VAMs. Despite concerns raised, large school districts such as Chicago, Houston, and New York City, as well as smaller school districts throughout the nation, have since adopted statistical modeling techniques to measure teacher effectiveness in similarly high-stakes ways (Amrein-Beardsley & Collins, 2012; Corcoran, 2010; Weisberg et al., 2009).

Methodological and Pragmatic Issues in Teacher Evaluation

Value-added models. The use of VAMs to measure teacher effectiveness is based on several theoretical and methodological assumptions about measuring the contribution of a teacher to the learning of his or her individual students. First and foremost, value-added estimates of teacher effects are treated as measures of “teacher effectiveness” (Berliner, 2014; Braun, 2005; Corcoran, 2010). In addition, it is assumed that the effectiveness of teachers is the most important variable in student achievement (Ballou, 2012; Sanders & Horn, 1998), an assumption challenged within the literature (Braun, 2005; Corcoran, 2010). Considerable research exists to suggest that other family and community variables strongly predict student achievement (Coleman et al., 1966; Rothstein, 2009, 2010). Specifically, the assumption that teachers who positively impact individual students also have the same effect on entire classes of students simplifies the complex interactions of numerous in- and out-of-classroom/school exogenous variables to a presumably one-directional relationship between teachers and their students (Berliner, 2014). Despite growing evidence that innumerable and often invisible variables

confound the attribution of a student's test score to his or her teacher (Berliner, 2014), proponents of VAMs argue that statistical controls using student background characteristics and prior achievement can account for family and community factors that are beyond the control of teachers and schools such that the models are useful accountability tools (Chetty, Friedman, & Rockoff, 2014; Harris, 2009).

Validity. Contemporary standards on test validity were set forth in the *Standards for Educational and Psychological Testing* (2014), sponsored by the American Educational Research Association (AERA), American Psychological Association (APA), and the National Council on Measurement in Education (NCME). This publication, the sixth edition to have been issued since 1954, was developed by an APA testing committee and reviewed by testing experts. Evidence of validity with regards to tests and their applications, defined by Messick (1980) as “the adequacy of a test as a measure of the characteristic it is interpreted to assess” (p. 1), has been established in the literature as paramount to the interpretation of test results and appropriate use of those interpretations to apply consequences.

VAMs purport to measure teacher effectiveness based on the assumption that a student's performance on a valid, reliable test measures his or her mastery of the aligned curriculum (Corcoran, 2010; Shavelson, Webb, & Burstein, 1986; see also Little, Goe, & Bell, 2009). The student's mastery is then attributed to teacher behaviors, again as a presumably valid, reliable measure of the teacher's effectiveness (Shavelson et al., 1986; see also Little et al., 2009). In order to evaluate a teacher evaluation system, it is critical, then, to examine different types of validity evidence with regards to the use of tests to make inferences about teacher quality (Herlihy et al., 2014).

The four most commonly gathered kinds of test validity evidence are content-related, criterion-related, construct-related, and consequential. Content-related evidence of validity suggests the extent to which a test measures the content, skills, or objectives that it is supposed to measure (i.e., the extent to which the test adequately samples the domain of content or behavior about which test results will be used to make inferences) (Messick, 1975; Popham, 1988). The use of teacher evaluation systems that rely on student test scores alone or in addition to classroom observations necessitates the gathering of content-related evidence supporting the standardized test results presumably as a measure of students' content or skill mastery (e.g., in reading or mathematics) (Fink, 1995; Herlihy et al., 2014).

Criterion-related evidence of validity often includes the correlation between performances on the measure of interest with an independent external criterion (Messick, 1975; Popham, 1988). In teacher evaluation systems, the use of an external criterion as evidence of validity such that the test (e.g., students' performance on AIMS reading as a measure of reading content mastery) and an external criterion (parent surveys with questions about the teachers' reading instructional skills) are not measures of the same domain of content or behavior is highly problematic. If student performance on standardized tests is a valid measure of teacher quality, states and districts implementing high-stakes teacher evaluation systems should be concerned if and when teachers' value-added scores are weakly correlated to an external criterion such as their classroom observational outcomes (Herlihy et al., 2014).

Construct-related evidence attempts to support validity differently such that multiple, varied types (a "network") of evidence are gathered to validate a test-based

inference (Popham, 1988, p. 123). Construct validation is most often conducted for hypothetical constructs such as “teacher quality” but can be used for attributes as well (e.g., reading or mathematics ability). Because test results are being used in teacher evaluation systems to make inferences about the quality of a teacher, an evidential basis for these uses is paramount. Messick (1975, 1980) argued that construct validity is important for test use as reliance upon criterion or content validity is insufficient. He further emphasized in later writings the conceptual need to describe validity as relevance in the context of use rather than as delineated types of evidence and to exercise caution in the interpretation and application of test results (Messick, 1980). He explained that evidence of validity must be based on the inferences drawn from the test results (Messick, 1980). Construct validation (evidential basis of test interpretation and use) should entail both confirmatory and disconfirmatory approaches such that convergent evidence supports the theoretical relationship between measures of the same construct and other variables and discriminant evidence that the measures are not related to exemplary measures of other distinct constructs (Messick, 1980, p. 1019).

Messick (1980) suggested that the implications of potential test uses (consequential validity of test interpretation and use) should be examined so as to “contrast the potential social consequences of the proposed testing with the potential social consequences of alternative procedures and even procedures antagonistic to testing” (p. 1020). In this way, test interpretation and use should be justified given the function or outcome of the test (either on an evidential or consequential basis) (Messick, 1980). In addition, Kane (2008) argued that the interpretation and use of test scores should be evaluated in context and then explicitly identified and described by test

developers and/or users (p. 81). Teacher evaluation systems that require the interpretation and use of test results to draw inferences about teacher quality should be supported by evidence of construct and consequential validity (Herlihy et al., 2014).

Despite confidence among VAM proponents, the validity of value-added estimates as measures of teacher effectiveness should be questioned for several reasons. Teachers' VAM scores have been shown to vary when using different test instruments (even within the same content area) (Bill & Melinda Gates Foundation, 2010; Corcoran, 2010; Lockwood et al., 2007; Papay, 2010; see also Darling-Hammond, Amrein-Beardsley, Haertel, & Rothstein, 2012). Lockwood et al. (2007) found large variation in the estimates of teacher effects between two subscales of a mathematics assessment relative to the variation between the various value-added models used in the study. Furthermore, the variation within teachers across the two assessments exceeded the variation across teachers (Lockwood et al., 2007). The Bill & Melinda Gates Foundation (2010) also found that teachers' estimated value-added varied across state and other standardized tests. Papay (2010) concluded as well that value-added scores varied based on the time of year in which the test was administered (e.g., early fall, mid-spring, or the end of the school year).

Briggs and Domingue (2011) also compared the value-added estimates produced by two different VAMs for teachers in LAUSD and found that only 46.4% and 60.8% of teachers would retain the same effectiveness ratings under both models for reading and mathematics outcomes, respectively. These findings are particularly concerning given that the study was conducted in response to *The LA Times'* decision to publish the names and value-added scores of LAUSD teachers, presumably as a means of holding

ineffective teachers publically accountable (Felch, Song, & Smith, 2010). The sensitivity of value-added estimates to different tests in the same content area and based on time of test administration raises concerns about measurement error and instruction that is narrowly focused on the test (Darling-Hammond et al., 2012).

Reliability. The reliability of a measure, meaning freedom from “measurement error” (Fink, 1995, p. 142), in any teacher evaluation system is paramount to the stability of the system. As a result of measurement error, obtained scores (e.g., students’ scores on AIMS, teachers’ observation scores) are different from true scores (only attainable through a perfect, error-free measure) (Fink, 1995). In order to be used as a reliable measure of teacher effectiveness, VAM estimates should be stable across years, courses, and statistical models (Baker et al., 2010; Newton et al., 2010).

While two recent studies have found moderate stability in value-added estimates over time (Koedel & Betts, 2007; McCaffrey, Sass, Lockwood, & Mihaly, 2009), other researchers have shown estimates of teacher effectiveness to vary considerably over time, across courses taught, and depending upon the statistical model used (Briggs & Domingue, 2011; Newton et al., 2010). Koedel and Betts (2007) ranked elementary mathematics teachers in San Diego and found that of those in the top and bottom quintiles, only 35% and 30%, respectively, remained in the same quintile across a two-year period. Notably, 31% of those teachers in the bottom quintile moved into the top two quintiles the following year (Koedel & Betts, 2007). McCaffrey, Sass, Lockwood, and Mihaly (2009) reported moderate correlations between teacher effectiveness ratings across years for elementary and middle school mathematics teachers in five large Florida school districts, indicating that only about one-third of the teachers who were ranked in

the top quintile one year remained in that quintile the following year, and approximately 10% of those top-ranked teachers moved to the bottom quintile (McCaffrey et al., 2009).

Newton, Darling-Hammond, Haertel, and Thomas (2010) also examined the stability of teacher rankings across years, courses, and models. While the researchers found that the rankings of more than half (56-80%, depending on the model used) and nearly three-quarters (74-93%, depending on the model used) of teachers changed one or more deciles from year to year, even greater variation existed among teachers depending on the courses they taught (Newton et al., 2010). Again, instability of VAM estimates across years, courses, etc. poses considerable challenges for schools and districts tasked with applying high-stakes consequences to teachers with low value-added scores (Berliner, 2014).

Fairness. Issues related to fairness must also be considered in any teacher evaluation system whether resulting from the instruments used or system design (AERA, APA, & NCME, 2014). When using students' test scores to estimate teacher effects, fairness or bias issues can arise regarding measurement quality, access to the constructs measured, and validity of individual score interpretation. With regards to teacher evaluation systems relying at least in part on VAMs, the implications of access and individual score interpretation are particularly relevant. Valid, reliable tests for non-core content areas such as social studies, art, music, physical education, etc. are not often available to assess student mastery. As a result, teachers of these subjects may be assigned a grade- or school-level value-added score based on student performance on standardized tests in reading and/or mathematics (as is the case in the district in this study). In order to make appropriate inferences, all individuals in the population (i.e.,

teachers across grade levels and content areas) must have fair access to the construct (i.e., equal and equitable opportunity to demonstrate professional quality) (AERA, APA, & NCME, 2014). Without valid, reliable tests aligned to the content taught by some teacher subgroups, schools and districts must justify the implementation of accountability systems that use VAMs to measure teacher effectiveness and apply high-stakes consequences.

With regards to design fairness, the non-random assignment of units/participants to treatment and control groups introduces possible bias. When conducting a quasi-experiment such that units have been non-randomly assigned, possible selection bias may be introduced as a result of preexisting groups (e.g., across or within schools) due to one or more characteristics of group members that are related to the treatment (Fink, 1995). The formation of nonequivalent comparison groups by any non-random sampling method (e.g., self-selection, purposive sampling, convenience sampling, etc.) threatens the validity of any treatment effect estimate (e.g., school, teacher, or program) (Rossi, Lipsey, & Freeman, 2004).

Non-random assignment of students into classrooms complicates the use of statistical models to estimate teacher effects. Studies have shown that random assignment of students within and between schools is far from the norm, especially at the elementary level (Braun, 2005; Burns & Mason, 1995; Harris, 2009; Monk, 1987; Paufler & Amrein-Beardsley, 2014; Player, 2010; Praisner, 2003), often resulting in the creation of classrooms with homogenous groups of students (e.g., from racial minority backgrounds, with limited English proficiency, receiving gifted and special education services, eligible for free or reduced lunches, etc.). Although observable variables are often used as

controls in VAMs to estimate teachers' value-added (Braun, 2005; Burns & Mason, 1995; Harris, 2009; Monk, 1987; Paufler & Amrein-Beardsley, 2014; Player, 2010; Praisner, 2003), researchers argue that bias may still be introduced into value-added estimates under these non-random conditions (Baker et al., 2010; Capitol Hill Briefing, 2011; Koedel & Betts, 2011; McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004; Rothstein, 2009, 2010).

Despite the inclusion of student-level variables in VAMs, most commonly background characteristics (e.g., racial or ethnic background, language proficiency, special education needs, socioeconomic status) and prior academic achievement (Harris, 2009; Sanders, 2006; Sanders, Wright, Rivers, & Leandro, 2009) as well as school-level variables (e.g., daily attendance rates, prior teachers' residual effects) (Sanders & Horn, 1998; Sanders & Rivers, 1996), the question remains as to whether these variables account for non-observable variables (e.g., parental support, access to summer school programs, tutoring and other supplementary resources) that could also potentially bias value-added estimates (Berliner, 2014; Rothstein, 2009, 2010). Thus far, researchers have demonstrated that the use of complex statistical controls has not mitigated the bias introduced by the non-random assignment of students to the extent that value-added estimates should be used to make high-stakes decisions about teacher tenure, promotion, and retention (Capitol Hill Briefing, 2011; Koedel & Betts, 2007; Rothstein, 2009, 2010).

Clinical Supervision. Traditional measures of professional quality based solely on teacher characteristics (e.g., education and credentials, experience) have largely been replaced by models comprised of professional practice and student achievement measures. As a result, evidence of validity, reliability, and fairness must also be

examined with regards to clinical supervision models based on classroom observations (i.e., the Danielson FFT).

Validity. The use of observation protocols to measure a theoretical construct such as teacher quality through specific performance-based, observable attributes merits somewhat different evidence of validity (Kane, 2001). According to Kane (2001, 2013), an *argument-based approach to validity* involves two parts: 1) a descriptive part with “network of inferences and assumptions leading from the scores to descriptive statements about individuals” and 2) a prescriptive part that requires making decisions based on the statements (p. 337). Kane (2001) argued that generating an observed score (e.g., on the Danielson FFT rubric) from a performance (e.g., instructional behavior) as an indicator of a theoretical construct (i.e., teacher quality) is based on the assumption that the score generalizes across sources of construct-irrelevant variance (e.g., lessons, raters, etc.) (p. 333).

The evidence of validity then should be based upon the proposed use of the observation scores. In terms of interpretation, the purpose of the observation should include evidence related to: 1) scoring (e.g., appropriate, consistently applied, bias free); 2) generalization (e.g., representative of lesson quality overall, has accounted for unexpected error); 3) extrapolation (e.g., scores on all lessons are related to quality that can be enacted, there is not systematic error); and 4) implications (e.g., appropriately associated with teaching performance, observed scores support implications) (Bell et al., 2012). The observation protocol used should be examined for evidence of validity related to each interpretation above.

Recent research has examined observation protocols such as the Danielson FFT for evidence of criterion-related validity. Triangulation as a validation strategy enhances the criterion-related validity or credibility argument through convergent or divergent findings related to teacher behavior and student test scores as measures of teacher quality (Popham, 1988). To this end, scholars have examined relationships between traditional measures of teacher quality most often considered for compensation, namely education credentials and experience; evaluations by principals; and seemingly more objective value-added measures. For example, Jacob and Lefgren (2005, 2008) conducted a study with elementary teachers in a mid-size school district in the western United States and questioned whether teacher characteristics (e.g., personality, education credentials, test scores, experience) should be used as measures of teacher productivity, an underlying premise of traditional single salary schedules. Harris and Sass (2009) conducted a similar study of 30 principals in a mid-size Florida school district, arguing that if teacher characteristics are weak indicators of productivity (Goldhaber, 2007; Hanushek, 1986, 1997) then other measures should be considered.

If VAMs are presumably valid measures of teacher quality as a construct, scholars can rightly expect teachers to post similar scores from year to year that also converge with other observational measures of teacher effectiveness (Amrein-Beardsley, 2008; Hill, Kapitula, & Umland, 2011). Jacob and Lefgren (2005, 2008) first compared seemingly subjective principal evaluations to teacher characteristics (i.e., education, experience, or actual compensation) and concluded that principals' evaluations were better predictors of future student achievement than these teacher characteristics. They then analyzed principals' ratings and teachers' value-added scores to determine whether

past principals' ratings and value-added scores could predict teachers' future value-added scores better than traditional measures focused on teacher characteristics (Jacob & Lefgren, 2005, 2008). Results indicated that while value-added measures are generally better predictors of future student achievement (i.e., the coefficient for teachers' prior value-added scores was nearly twice as large as that for principal ratings), the measures are comparable in their ability to identify the best and worst teachers (Jacob & Lefgren, 2005, p. 30). However, findings also demonstrated that the principals' overall ratings were far better predictors of future parent requests (i.e., as a measure of satisfaction) for individual teachers than both teacher characteristics and value-added scores (Jacob & Lefgren, 2005, p. 30).

Harris and Sass (2009) conducted similar analyses, and findings demonstrated that teachers' value-added scores were at best weakly correlated with teacher characteristics such as education credentials and experience and that principals' evaluation ratings appeared to consider teacher characteristics, namely professional expertise (i.e., education, experience, and content knowledge) and personality (Harris & Sass, 2009). In addition, when value-added scores were calculated using only one year of test data, principals' evaluations better predicted future teacher value-added scores than did past scores. Notably, the use of multiple years of test data in value-added estimates improved their predictive capacity (Harris & Sass, 2009). However, even when using multiple years of data, principals' ratings still added information to (i.e., accounted for variance in) predictions of teachers' value-added scores in reading, although somewhat less so in mathematics (Harris & Sass, 2009). The results from these studies suggest that principals' evaluations are particularly important for determining the productivity of

novice teachers with few years of test score data (Harris & Sass, 2009); however, it is important to note that in general principals may struggle to distinguish between teachers of average quality as indicated by their value-added scores (Jacob & Lefgren, 2005, 2008).

Sartain, Stoelinga, and Brown (2011) examined evidence of criterion-related validity for the Danielson FFT based on its use in Chicago Public Schools. The researchers explored the relationship between observation ratings and student achievement to determine whether “teachers who receive higher ratings also tend to have students who achieve greater test score growth” (Sartain, Stoelinga, & Brown, 2011). In total, 501 teachers of English language arts and/or mathematics in grades 4-8 with both ratings on the Danielson FFT (i.e., Unsatisfactory, Basic, Proficient, or Distinguished) and a value-added indicator participated in the study. Collectively, this sample of teachers had 955 principal observations and on average were rated “Proficient” on the Danielson FFT. Sartain et al. (2011) concluded based on this sample that teachers’ observational ratings for the five components each in Domains 2 and 3 (i.e., Creating an Environment for Learning and Teaching for Learning, respectively) were statistically significant predictors of their value-added scores such that teachers with the lowest or highest observation ratings also tended to have the lowest or highest value-added scores in both reading and mathematics. However, the researchers also acknowledged limitations in the student data systems, especially at the elementary level, that made the attribution of student achievement growth to one teacher difficult (e.g., team teaching and other arrangements) and noted that many teachers in the district did not teach tested subjects or grade levels (Sartain et al., 2011).

Tyler, Taylor, Kane, and Wooten (2010) had previously conducted a similar study in the Cincinnati Public School system, specifically measuring the relationship between Domains 2 and 3 on the Danielson FFT and student growth scores. They argued that an increase of one point in teachers' overall average Danielson scores was associated with a student achievement gain in mathematics (one-sixth of a standard deviation) and reading (one-fifth of a standard deviation) (Tyler, Taylor, Kane, & Wooten, 2010). However, the researchers also acknowledged that their findings only measured teacher practices in those two domains and could not determine the relationship of other practices to student achievement growth (Tyler et al., 2010).

While these studies measure the relationship between the Danielson FFT and student growth scores as evidence of criterion-related validity (Sartain et al., 2011; Tyler et al., 2010), research examining evidence of construct-related validity with regards to the use of the Danielson FFT to measure teacher quality is very limited (Sloat, 2014). Additional research in this area, specifically focused on the Danielson FFT and other professional practice rubrics, is needed given the high-stakes consequences associated with many federally-supported, state policy-directed teacher accountability systems now in use.

Reliability. Observations of teacher behavior are subject to measurement error, as well, both from the instruments used and personnel conducting the observations. According to the Bill & Melinda Gates Foundation (2012), a reliable observation instrument should consistently reflect a teacher's instructional quality, meaning that the teacher's rating "should be due to the quality of the lesson and not the quality of the observer" (p. 17). To this end, the Bill & Melinda Gates Foundation (2012) sought to

examine the validity and reliability of five observational instruments including the Danielson FFT. The study involved 900 trained raters who observed 7,491 videos of 1,333 teachers delivering instruction in grades 4-8 across six districts (Bill & Melinda Gates Foundation, 2012). With regards to the reliability of the instruments, findings demonstrated that a relatively small proportion of the variation in overall scores among teachers (14% to 37%, respectively) was due to consistent differences (meaning that an observation score from any single lesson was affected by inconsistent aspects of a teacher's professional practice) (Bill & Melinda Gates Foundation, 2012, p. 17). Although the course section (i.e., students in the classroom) accounted for less than 5% of the variation in overall scores, an individual teacher's variance in scores across lessons was "at least half as large as the teacher effect" described above, meaning that his or her score for any single lesson would not represent overall instructional quality (Bill & Melinda Gates Foundation, 2012, p. 17-18). Although 10% or less of the variance in total scores was attributable to rater inconsistencies, raters often scored individual lessons differently (Bill & Melinda Gates Foundation, 2012, p. 18). This finding is particularly important given the reliance of many teacher evaluation systems on a single classroom observation for any given teacher.

Acknowledging that inaccurate classroom observations undermine trust and negatively impact decision making, the Bill & Melinda Gates Foundation (2012) suggested that observations of more than one full lesson by more than one well-trained observer should be conducted to reduce error and increase inter-rater reliability. Although these implementation measures are critical when teacher evaluation results are associated

with high-stakes consequences, schools and districts may have inadequate time and resources to ensure multiplicity of observations and raters.

Ho and Kane (2013) also examined different combinations of 129 observers and video-taped lessons for 67 teachers to evaluate the accuracy and reliability of the observers (all school-site personnel) using the Danielson FFT. Each observer scored 24 lessons, providing more than 3,000 video scores for the study (Ho & Kane, 2013). The researchers noted that observers rarely rated the teachers' instruction in the top or bottom categories (Unsatisfactory or Distinguished) on the Danielson FFT (i.e., central tendency error) (Ho & Kane, 2013). Administrators differentiated more among teachers in terms of scores than did teacher peers and also rated teachers from their own schools higher than both administrators from other schools and teacher peers (Ho & Kane, 2013). Teachers also rated themselves more favorably than peers; however, this phenomenon did not affect the relative ranking of teachers (Ho & Kane, 2013). In addition, a positive (or negative) impression of a teacher formed in an early video often lingered throughout subsequent videos (Ho & Kane, 2013). Finally, and perhaps unsurprisingly, having more observers rate each video increased the reliability of the ratings (Ho & Kane, 2013). Again, these findings are particularly relevant for validity and reliability given that teacher evaluation systems often require only one or two observations by a single or few administrators/evaluators over the course of the school year.

Sartain et al. (2011) found similar results in their examination of the use of the Danielson FFT in Chicago Public Schools. With regards to the reliability of observations by principals and external observers, findings demonstrated that principals reliably rated teaching practice at the low and middle levels of the scale (Sartain et al., 2011). It is

important to note that principals were also more likely to rate a teacher as “Distinguished” when external observers rated the same individual as “Proficient” (Sartain et al., 2011). While most principals agreed with the external observers on ratings of teaching practice overall, 11% and 17% of principals gave consistently lower and higher ratings than did the observers, respectively (Sartain et al., 2011). These results reinforce the need to examine the reliability of instruments given that any instrument, especially an observation protocol when used in a single lesson observation by only one observer, can and under such conditions would likely yield unreliable results. Although states and districts should implement evaluation systems with components that will likely increase reliability (e.g., multiple raters or observers, multiple observations, inter-rater reliability checks), these efforts pose practical challenges such as discerning an optimal number of lessons and raters needed to adequately generate reliable scores (Herlihy et al., 2014).

Fairness. Proponents of VAMs argue that there are several practical advantages to using statistical models to measure teacher effectiveness over the more subjective judgments of an observer (Little et al., 2009). Much research has been conducted to better understand principals’ beliefs and perceptions, and the impact of these on evaluation processes. Weisberg, Sexton, Mulhern, and Keeling (2009) examined observational measures of teacher effectiveness in one such study involving 15,000 teachers, 1,300 administrators, and 80 local and state education officials in 12 districts across four states. Based on study results, they argued that school districts falsely assumed that teacher effectiveness across classrooms looks and sounds the same—a phenomenon referred to as the *Widget Effect* (Weisberg et al., 2009). As a result, individual teachers were not

recognized as professionals with unique strengths and weaknesses but instead as interchangeable parts that can be removed and replaced with seemingly little impact on students' learning (Weisberg et al., 2009, p. 4).

Weisberg et al. (2009) identified several contributing factors, citing that evaluations: 1) were brief and sporadic (typically two or fewer per year); 2) were often conducted by evaluators who had not been provided with adequate training; and 3) effectively ignored variation in instructional effectiveness. In fact, the vast majority of teachers observed in the study received "Satisfactory" ratings on a binary scale or one of the two highest ratings when multiple ratings were available (Weisberg et al., 2009). The failure to differentiate based on performance fostered an expectation among teachers that they would receive a high rating (Weisberg et al., 2009).

Weisberg et al. (2009) concluded that these teacher evaluation systems were intended at best to capture a snapshot of instructional performance rather than actually differentiate between teachers based on their impact on student learning and achievement. In turn, excellent teachers did not receive recognition or compensation for their performance, and poor teachers were not identified for additional support and professional development (Weisberg et al., 2009). The small percentage of teachers identified as needing improvement were observed on average less than three times per year and received the same limited amount of specific feedback as their colleagues who received higher ratings (Weisberg et al., 2009, p. 20-21). In addition, novice teachers as a group were neglected in that they did not receive frequent, meaningful feedback, or adequate support in the early years as they developed their skills (Weisberg et al., 2009).

Other studies have examined principals' perceptions regarding the alignment between value-added estimates of teacher effectiveness and their own formal evaluations based upon classroom observations and other indicators of professional capacity and performance (e.g., evidence of instructional planning, leadership, reflection, etc.). Childers (2012) conducted a qualitative study examining the alignment of methods to determine the extent to which value-added estimates as well as principals' evaluations measure teacher quality and under what conditions (time and place) each method is most appropriate. Through a case study approach, she examined the use of new observation protocols and value-added data as part of the implementation of the SAS® Education Value-Added Assessment System (EVAAS®) in a large urban school district in North Carolina (Childers, 2012).

Specifically, Childers (2012) questioned administrators about their perceptions of the alignment between value-added estimates and their own knowledge about and observations of teacher performance in their schools and asserted that they held several common beliefs about teacher quality: 1) high quality teachers can be distinguished from good teachers in directly observable ways (p. 20-21); 2) evaluation protocols can be used to monitor teacher performance and provide teachers with performance guidelines (p. 25); 3) value-added data can be used to confirm principals' direct observations of teachers' content knowledge and student learning and should facilitate discussions about improving professional practice (p. 28); and 4) principals generally perceive formal evaluations based on observations and value-added data to be aligned but believe that the methods should be used "in tandem as confirmatory evidence" as each measures different aspects of teacher quality (p. 30-31). She noted that principals' perceptions varied on an

individual basis, reflecting a need for increased dialogue about the alignment of quantitative and qualitative measures of teacher quality (Childers, 2012).

Designing, Implementing, and Improving Teacher Evaluation

Several common themes exist in the literature for designing, implementing, and improving teacher evaluation systems. For example, Toch and Rothman (2008) identified four key components of a comprehensive evaluation system, emphasizing the need for: 1) clear, specific standards with a scoring rubric; 2) multiple measures of effectiveness (e.g., observations, student work, evidence of collaboration with parents, etc.); 3) multiple evaluations by multiple evaluators; and 4) frequent, meaningful feedback and appropriate professional development. In addition, school administrators should be empowered to make decisions about compensation and retention based on the results of an evaluation system meeting these guidelines (Toch & Rothman, 2008).

Weisberg et al. (2009) made similar recommendations with regards to the development of teacher evaluation systems including: 1) differentiation between teachers based on their effectiveness in promoting student achievement (e.g., clear standards, multiple rating options, regular monitoring, meaningful feedback, targeted professional development, and intensive support); 2) comprehensive training and accountability for evaluators; 3) use of evaluation results in the decision-making process (e.g., assignment, compensation, retention, and dismissal); and 4) lower-stakes exit options with a fair, efficient due process system for teachers who wish to leave the district (p. 30).

Tucker and Stronge (2005) cited the need for similar components in teacher evaluation systems; however, they also emphasized the need to use student achievement to measure teacher effectiveness. Based on their examinations of systems in four districts,

Tucker and Stronge (2005) advocated for evaluation systems that utilized student gains in conjunction with observations:

Given the clear and undeniable link that exists between teacher effectiveness and student learning, we support the use of student achievement information in teacher assessment. Student achievement can, and indeed should be, an important source of feedback on the effectiveness of schools, administrators, and teachers.

(p. 102)

This recommendation in the literature has been increasingly applied in practice.

As multiple measures have been championed as a means to more validly and reliably represent the construct of teacher quality (AERA, 2000; Baker, 2003), teacher evaluation systems, increasingly comprised of measures including classroom observations and student growth models, have been widely adopted across the nation. Despite widespread support for such evaluation systems among many policymakers, some educational researchers and scholars, and the general public (Bill & Melinda Gates Foundation, 2013; Tucker & Stronge, 2005), methodological and pragmatic issues persist among these models.

Using Teacher Evaluation Systems for High-Stakes Decisions

Some VAM supporters argue that these metrics should be used by principals to make human capital decisions in their schools. For example, Jacob (2012) argued that a teacher's contribution to student learning is the "most meaningful measure of teacher quality" (p. 11) and that given the inconsistencies in predictions based on observable characteristics (e.g., experience, education and credentials, etc.), VAMs may be the most useful metrics for attracting and developing talented teachers. By using VAMs to identify

areas for professional development, Jacob (2012) argued that principals could match high quality mentor teachers to those with specific needs for improvement. Furthermore, she suggested that principals should not only use value-added measures as a significant part of the teacher evaluation process but also as a recruitment, assignment, development, and retention tool in their role as human capital managers (Jacob, 2012).

If this is the case, states and school districts certainly have a vested interest in the validity, reliability, and fairness of the measures used in their new teacher evaluation systems (Herlihy et al., 2014). A myriad of intended and unintended consequences associated with perceivably invalid, unreliable, and biased measures have already posed significant legal and staffing problems for some districts and will likely continue to do so (Amrein-Beardsley & Collins, 2012). As a result, schools and districts that have designed and implemented systems perceived as such but lack adequate professional development to support those who receive lower than desirable scores may not only suffer a decline in morale among current staff but also struggle to recruit, hire, and retain teachers in the future, especially those arguably savvy enough to “shop” for schools with seemingly less rigorous evaluation systems (Herlihy et al., 2014, p. 7-8). Certainly, a lack of validity, reliability, and fairness evidence for the measures increasingly used in teacher evaluation systems, especially given the variability of systems across states and districts, should deter their use for high-stakes employment decisions related to teacher pay, tenure, or termination (Konstantopoulos, 2014).

The development and implementation of teacher evaluation systems with high-stakes consequences necessitates further examination of the perceptions of school administrators and teachers as the subjects of these systems in practice. In Chapter 3, I

describe in greater detail the conceptual framework I applied in this study. These concepts formed the platform for the research questions and design discussed in Chapter 4. Based on study results presented in Chapter 5, I also evaluate the utility of these concepts for understanding stakeholder perceptions in context in Chapter 6.

CHAPTER 3

Conceptual Framework

Much research has been conducted in recent decades examining school reform including the purpose, implementation, and effectiveness of micro-level reforms in school contexts as well as reforming schools as a macro-level process. Arguably, most reformers intend to improve public schools for all students, especially those at greatest risk of failure, regardless of whether they hold positions at the state or federal levels of government; in universities, schools of education, or other public institutions; at foundations, nonprofits, or corporations; or at the local level in districts and schools (David & Cuban, 2010, p. 181). They are confident that the identified problems are both real and complex, but if solved, the results would enhance the functions of schooling through increased student learning and achievement (David & Cuban, 2010). Yet, despite high hopes, carefully laid plans, and best efforts, many reforms are unpackaged quite differently at the school and classroom levels.

School Reform as Embedded in Contexts

Schooling has also been situated as a social, cultural, and historical institution defined and redefined by reforms within the institution (Popkewitz, 1991, p. 13). Rather than describing school reform as a formal process, Popkewitz (1991) framed schooling and school reform from a postmodern paradigm as a narrative that must include understandings about knowledge (epistemology), power, and institutions. He argued that expert systems of knowledge shape human thought and actions in a presumably commonsensical way (Popkewitz, 1991, p. 5). However, these systems of knowledge while presupposed to be natural are not natural at all. Instead, they are ideas that shape

personal decisions and even understandings about what choices are possible, logical, and rational (Popkewitz, 1991). In this capacity, expert systems of knowledge in education have a powerful effect, directing the way that teachers think, feel, and talk about their practice, children, and learning (Popkewitz, 1991, p. 5).

School reform occurs within the confines of constructed understandings about the purpose of schooling, de facto assumptions about the rationality of expert knowledge about schooling, and the definition of professionalism and practice within schools as institutions (Popkewitz et al., 1982). Characterizing the term “reform” as an evolving concept that has embodied different meanings over time, Popkewitz (1991; see also Popkewitz, 2008) suggested that current uses of the term vary based on one’s particular view on individualism and understanding of professional practice—the term assumes different meanings depending on the context in which it is used (p. 14).

School Reform as a Policy Cycle

Juxtaposing the notions of perpetual educational evolution as “progress” and educational reform as “cyclical,” Tyack and Cuban (1995) argued that in reality both cases can occur in tandem. They suggested that it is entirely possible for policy talk to cycle even when institutional trends have not (Tyack & Cuban, 1995). Specifically, they described educational reform as occurring in cycles, defining policy talk, the first phase of reform, as the “diagnosis of problems and advocacy of solutions” (Tyack & Cuban, 1995, p. 40). During the next phase of policy action, Tyack and Cuban (1995) explained that reforms are adopted through legislation, school board regulations, or the decisions made by others in positions of authority (p. 40). Lastly, the actual implementation of “planned change in schools [by] putting reforms into practice is...often much slower and

more complex than the first two [phases]" (Tyack & Cuban, 1995, p. 40). Although many within education have observed and criticized seemingly repetitious calls for the same or very similar reforms, Tyack and Cuban (1995) argued that this policy talk occurs in different contexts over time as steady, albeit slow changes in schools as educational institutions slightly reframe the conversation. Comparing this changing dialogue to a swinging pendulum, Cuban (2008) attributed much of the swing to competing public social values, despite a shared faith in schools as the catalyst for collective and individual improvement. This staunch belief in schools as the solution to imminent economic or social crises remains in the face of shifting public attention and little consensus as to what constitutes a crisis (Cuban, 2008).

Challenges in directly linking changes or trends in schools to policy talk are attributable to three major factors: the time lag between advocacy and implementation, the uneven penetration of reforms across schools, and the varying impact of reforms on social groups (Tyack & Cuban, 1995, p. 55). According to Tyack and Cuban (1995), election deadlines, career advancement opportunities, the availability of grant funds, and shifts in media attention prompt reformers to redefine or decide to ignore problems for which they once sought solutions. Even so, the implementation of now seemingly obsolete solutions may be underway if not already completed. Furthermore, variations in school systems across the nation are often disregarded in the reform dialogue. Unsurprisingly, reforms are implemented at different times and at varying rates depending on the location, character, and demographics of each school community (Tyack & Cuban, 1995, p. 56).

The reforms most likely to last long enough to shape institutional trends share certain commonalities. These reforms are often structural add-ons intended to enhance rather than disturb school processes and procedures (Tyack & Cuban, 1995, p. 57). Similarly, reforms that are non-controversial or those supported by influential constituencies are most likely to receive adequate support (Tyack & Cuban, 1995). Lastly, those required by law and easily monitored are most likely to be implemented, although most often as a result of enforcement mechanisms or incentives rather than overwhelming public support (Tyack & Cuban, 1995). Tyack and Cuban (1995) aptly summarized the connection between policy talk and institutional trends, concluding that in reality “whether policy talk led to implementation depended much on who was talking” (p. 58).

Defining what it means to be a “good” school has driven much of the policy talk in education for decades (Cuban, 2003). Conflicting views of schools as “a virtual arm of the economy” (Cuban, 2003, p. 53) as opposed to centers for “building literate and moral citizens committed to democratic equality” (p. 41) have prompted much of the vacillation in reform efforts. Cuban (2003) argued that this fundamental disagreement about the purpose of schools has only been exacerbated by inevitable changes in the economic, social, political, and demographic American landscape. In recent decades, those supporting the standards-based, test-driven accountability movement have also dominated the policy conversation about how to ensure that all students attend a “good” school (Cuban, 2003). The use of teacher- and school-focused accountability systems to both identify and reward “good” schools and teachers has become the hallmark of 21st century education reform.

Measuring the Effectiveness of School Reform

Evaluating the effectiveness of school reform also depends on whose standards of measurement are used. In order to determine whether an innovation has been successful, one must ask how success is to be measured. Cuban (1998; see also Tyack & Cuban, 1995) argued that the standards set by policymakers, administrators, and researchers often relate to whether the goals of a program were achieved, to what extent the program was popular, and whether the program was implemented with fidelity (p. 456-458). These standards typically rely on quantitative results (e.g., students' standardized test scores) to determine success. Cuban (1998) described the origins of the first standard, referred to as the *effectiveness standard* and related to program goals, noting that:

For the last quarter-century the effectiveness standard, an outgrowth of a strong belief in professional expertise and technical rationality applied to organizations, has been used for schools to examine what students have learned in school and do after graduation by using proxy measures for both such as student test scores....
(p. 456)

Cuban (1998) further explained that the desired goals for reform and their subjective measures of success are determined by those in positions of authority (e.g., national and state policymakers).

In addition, Cuban (1998) cited *popularity* as the second standard by which policymakers and others in positions of authority often measure success. He noted that the perception of “fashionableness” is particularly important as a prerequisite for support among many public officials as evidenced by their careful attention to public opinion polls and media reports (Cuban, 1998, p. 457). If reforms seem to be popular amongst

constituents, then policymakers are less likely to balk at expenditures of public funds to address seemingly urgent problems facing schools.

Third, the *fidelity standard*, an important component in measuring effectiveness overall, relies heavily on implementers (e.g., administrators and teachers) who are tasked with following the program blueprint (Cuban, 1998, p. 458). Cuban (1998) defined fidelity as a measure of “the fit between the initial design, the formal policy, the subsequent program it spawns, and its implementation” (p. 458). In order for a program to be deemed effective across contexts, implementers in each context must adhere to the original design when putting the program into practice (Cuban, 1998). However, the use of these three standards by policymakers, administrators, researchers, and others to measure effectiveness and in essence legitimize the reform do not necessarily reflect the standards by which practitioners make the same judgments.

Practitioners often have divergent criteria for measuring effectiveness. Their vantage point as “the foot-soldiers of every reform aimed at improving student outcomes” prompts the use of an entirely different set of standards (Cuban, 1998, p. 459). While teachers certainly seek to improve student performance and attitudes, students’ standardized test scores are seldom the measuring tool used. Rather, teachers consider students’ attitudes, values, and behaviors on both academic and nonacademic tasks in various contexts as indicators of their learning (Cuban, 1998, p. 459). Teachers often seek to alter and adapt reforms during implementation—these actions are seen as “healthy signs of inventiveness, active problem solving, and a precondition for determining effectiveness” by their own standards (Cuban, 1998, p. 460). In fact, this *standard of adaptiveness* is considered essential in order for a reform to meet the other most important

standard for practitioners—that of *longevity* (Cuban, 1998). In order for a reform to be considered a success to most teachers, the reform must outlast the next cycle of policy talk.

Symbolic Adaptation of School Reform

Measuring the success of educational reform may be more easily understood as having multiple layers of meaning. According to Popkewitz et al. (1982), the “publically accepted criteria or standards by which people judge success or failure” may constitute a *surface* layer of meaning (p. 9). However, the *underlying* layer would include “the socially accepted procedures, guidelines, and assumptions...that make the activities, interactions, and teaching/learning experiences in institutions seem plausible and legitimate” (Popkewitz et al., 1982, p. 9). In this regard, the underlying meaning often supports and reinforces the surface meaning. School reform is most commonly evaluated in terms of its efficiency in meeting the criteria or standards at the surface; however, these measures rarely account for the modification of content and culture that inherently occurs through schooling, the biases and selection that occur in the culture transmission process, and the relationship between school practices and social commitment that is often hidden behind or obscured by rituals, ceremonies, and slogans (Popkewitz et al., 1982, p. 11).

In an effort to better understand teacher evaluation systems as school reform, the way that the reform responds to different school contexts must be considered. Power in schooling shapes the ways that individuals construct their identities and understand their experiences over time (Popkewitz, 1991, p. 14). From this perspective, “the act of reform, in contrast to mere change, is an act of social commitment and reaffirmation” of the

ideals individuals associate with schooling (Popkewitz et al., 1982, p. 3). Channeled reform then not only reinforces but also legitimizes existing social values, especially concerning authority and control in schools (Popkewitz et al., 1982, p. 5). Popkewitz et al. (1982) found that an important component of the conceptualization of knowledge, namely the *professional ideology* at the school and within the community that guided the behavior of administrators and teachers, profoundly impacted the implementation of the reform. This professional ideology was regulated and reinforced by the school districts, state departments of education, and communities.

In the case of teacher evaluation systems, this reform can be seen as a mechanism by which professionals can “demonstrate publically their efficiency in meeting certain goals of public education” (Popkewitz et al., 1982, p. 165). As such, teacher evaluation legitimizes the teacher as a professional and acts as a symbol of credibility for the institution as having met its social mandate (Popkewitz et al., 1982, p. 169). The use of observations to supervise teachers exemplifies a ritual or ceremony that further supports the public image of schooling as legitimate and teachers as rational professionals. To some extent, the evaluation process could arguably develop among teachers a commitment to the occupation of educator and to the stability of schooling as an institution; however, these do not preclude school administrators or teachers from participating in or adapting reform activities (i.e., in this case evaluation processes and procedures) in a ceremonial or symbolic way.

Based on this argument, reform serves to legitimize schooling as an institution and to some extent protect the same institution from public scrutiny. Reform symbols, slogans, and rituals reflect “potential actions” but are not necessarily descriptive of what

is actually happening or the motives of those involved (Popkewitz et al., 1982, p. 20-21). In this way, teacher evaluation systems may in reality “conserve rather than change” (Popkewitz et al., 1982, p. 21) procedures, rules, and practices in schools through symbolic action that is unreflective of the real activities of supposed reformers.

Understanding Stakeholder Perceptions

The application of this conceptual framework, specifically school reform as a policy cycle, to examine a state policy-directed, locally-developed teacher evaluation system acknowledges and reflects the diverging views held by various stakeholders including policymakers, researchers, practitioners, parents, and the general public. The use of the policy cycle as a conceptual perspective through which to understand teacher evaluation at the district level offered a means of situating a micro-level process within a macro-level structure. According to this perspective, stakeholder views reflect the current policy talk that drives the design and implementation of teacher evaluation systems based at least in part on quantifiable student performance outcomes and also determine the effectiveness standards to which these systems will be held—standards that will no doubt vary as well depending upon who is asked to take the measurements.

The language used by Cuban (1998) to describe various standards of effectiveness as defined by different stakeholder groups (i.e., effectiveness/purpose, fidelity of implementation, popularity, adaptiveness, longevity) has been useful; however, I also recognized the need to situate these standards of effectiveness within socially constructed understandings of the purpose of schooling, professional ideology of educators, and symbolic nature of actions taken in the name of school reform (Popkewitz, 1991). I used both of these conceptual perspectives in conjunction to better understand stakeholder

perceptions with regards the design, implementation, and evaluation of a teacher evaluation system in the district in this study.

In the next chapter, I describe the sequential mixed methods research design used in this study. I also discuss processes for instrument development, data collection, management, and analyses. Additionally, I outline the research activities I completed to establish the validity, reliability, and generalizability of results and acknowledge the study limitations.

CHAPTER 4

Methods

Role of the Researcher

District administrators in this study expressed a keen interest in better understanding stakeholder perceptions regarding the new teacher evaluation system that had, at the point of this study, been recently adopted. As a doctoral student and researcher, I was in a unique position to propose this study as a means of examining their concerns related to system development and implementation. I conceptualized and conducted the study with their full support as part of a larger, comprehensive evaluation that included as participants high school administrators and teachers as well as other certified staff members. The research design, data collection processes, and analyses described in this chapter reflect a subset of the comprehensive research questions and activities included in the larger district-level evaluation. I am grateful for the assistance of district administrators and staff who, therefore, assisted me throughout this process. Having assumed primary responsibility for this study, though, I independently developed the research design and conceptual framework, collected and analyzed data, and generated the findings and conclusions described in this and the forthcoming chapter.

Pragmatic Paradigm Stance

Researchers have long debated the paradigmatic differences between traditional quantitative and qualitative research, and there is certainly merit in carefully considering the inherent strengths and weaknesses in each (Johnson & Onwuegbuzie, 2004). However, Patton (1990) challenges those who ascribe to a singular paradigm or worldview by taking a more pragmatic stance to research, arguing that “the quality of a

study [should be judged by] its intended purposes, available resources, procedures followed, and results obtained, all within a particular context and for a specific audience” (p. 71-72). Johnson and Onwuegbuzie (2004) concurred, arguing:

[Pragmatism offers] an immediate and useful middle position philosophically and methodologically..., a practical and outcome-orientated method of inquiry that is based on action and leads, iteratively, to future action and elimination of doubt; and...a method for selecting methodological mixes that can help researchers better answer many of their research questions. (p. 17)

Further, Johnson and Onwuegbuzie (2004) cited the value of mixed methods research as “inclusive, pluralistic, and complementary” in that it allows “researchers [to] take an eclectic approach to methods selection” which should align to research questions in a logical and useful way (p. 17). In conceptualizing the research design for this study, I therefore embraced a pragmatic stance as an alternative paradigm (Greene, 2007), seeking to select methods based on their utility in answering my research questions and recognizing their complementary strengths and nonoverlapping weaknesses (Johnson & Turner, 2003; see also Johnson and Onwuegbuzie, 2004).

Mixed Methods Research Design

Johnson and Onwuegbuzie (2004) defined mixed methods research “as the class of research where the researcher mixes or combines quantitative and qualitative research techniques, methods, approaches, concepts or language into a single study” (p. 17). In an effort to better understand the perceptions of as many participants as possible with regards to different facets of the same complex phenomenon (Greene, 2007)—in this case, a state policy-directed, locally-developed teacher evaluation system—I used a

sequential mixed methods design with two stages or phases (Creswell, Plano Clark, Gutmann, & Hanson, 2003; Tashakkori & Teddlie, 2003; Teddlie & Tashakkori, 2006). The multistrand design (i.e., multiple phases such that each encompasses conceptualization, experiential [methods/analyses], and inferential stages) permitted the initial use of qualitative methods (i.e., school administrator and teacher interviews) to collect and analyze data that informed the subsequent development of quantitative methods (i.e., respective surveys) in the next phase (Teddlie & Tashakkori, 2006). See my research design model illustrated in Figure 2.

Integration of methods occurred during multiple stages of inquiry: 1) experiential – methods (i.e., transformation of qualitative themes into Likert-type survey items, inclusion of open-ended questions on the survey instrument); 2) experiential – analysis (i.e., transfer of quantitative demographic survey data into a qualitative database to analyze subgroup responses); and 3) inferential (i.e., triangulation of qualitative and quantitative findings for convergence) (Creswell et al., 2003, p. 173; see also Greene, 2007). By mixing methods for the purpose of triangulation, I sought to harness the strengths of different methods, seeking evidence of convergence and divergence of qualitative and quantitative data (Greene, 2007). In this regard, the status of the methods was equal such that neither the qualitative nor quantitative methods were dominant (Teddlie & Tashakkori, 2006; see also Creswell et al., 2003). The mixed methods design allowed me to capitalize on the inherent value of a qualitative method to “represent [a] social phenomena textually” and quantitative method to “represent [a] social phenomena numerically” (Greene, 2007, p. 99), serving as a means to “elaborate, enhance, deepen, and broaden the overall interpretations and inferences” drawn from this study (p. 101).

Research Design

Sequential Mixed Methods (Multistrand)

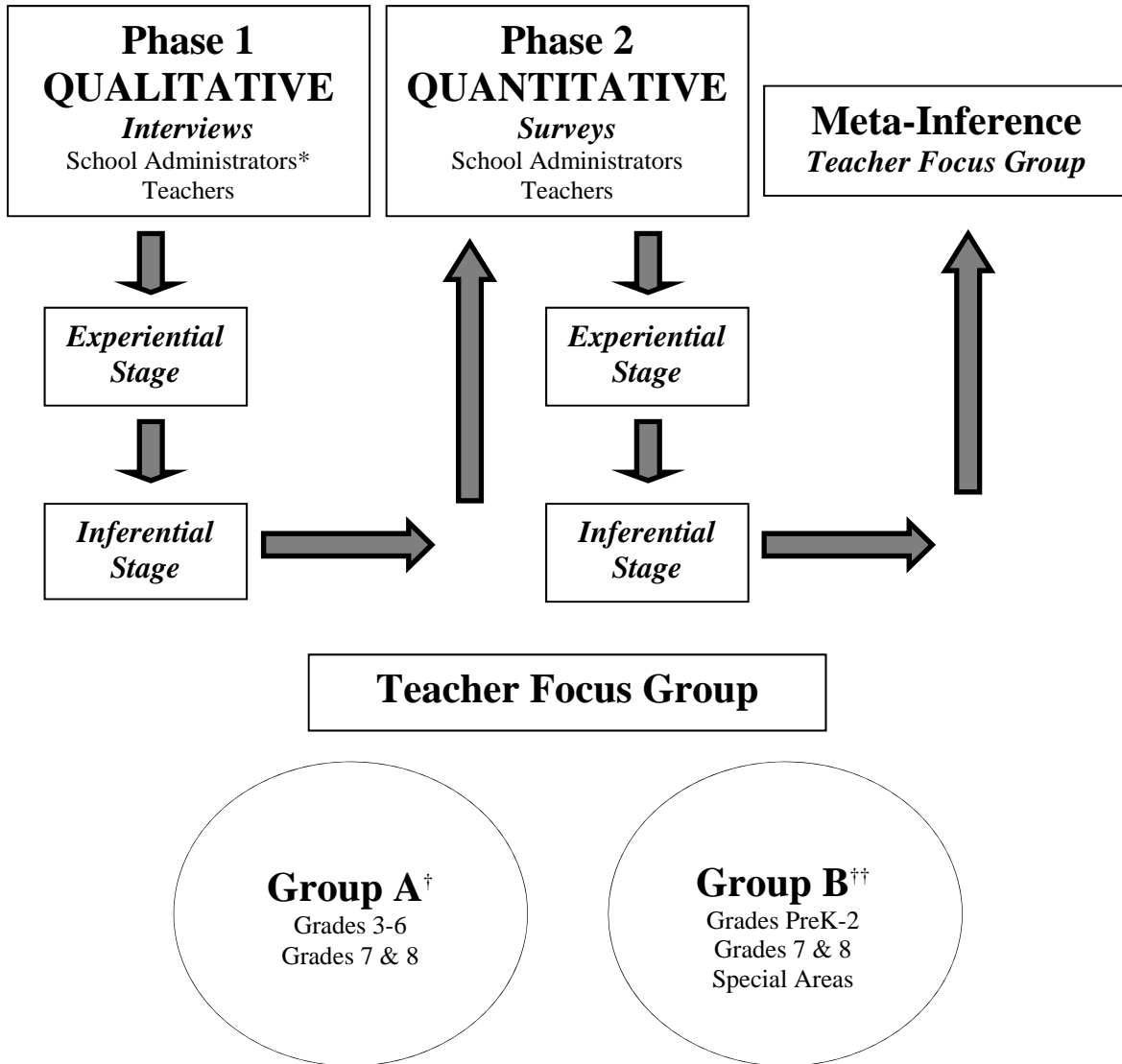


Figure 2. Sequential Mixed Methods Research Design. Adapted from “A General Typology of Research Designs Featuring Mixed Methods,” by C. Teddlie and A. Tashakkori, *Research in the Schools*, 13(1), p. 22. Reprinted from *Research in the Schools*, Copyright 2006 by the Mid-South Educational Research Association, Nashville, Tennessee. Reprinted with permission of the original copyright holder (see Appendix C).

* Principals and assistant principals.

† Grades 7-8 include English language arts, mathematics, and/or science (grade 8).

†† Grades 7-8 include social studies and science (grade 7). Special areas include art, music, and physical education.

In an effort to determine the applicability of the conceptual framework, specifically whether the standards of effectiveness conceptualized by Cuban (1998) (i.e., purpose, fidelity of implementation, popularity, adaptiveness, and longevity) were useful in understanding the perceptions of participants in this context, I used a sequential mixed methods research design to iteratively develop instruments and collect data in two phases (Creswell et al., 2003; Tashakkori & Teddlie, 2003; Teddlie & Tashakkori, 2006). Data analyses and findings in each phase informed subsequent research activities (Creswell et al., 2003; Greene, 2007; Tashakkori & Teddlie, 2003; Teddlie & Tashakkori, 2006). In the next section, I describe in greater detail processes for participant selection as well as data collection and analyses in each phase. I also discuss limitations in this study.

The population for this study included all 38 elementary school administrators ($n = 20$ principals including one serving as a substitute and $n = 18$ assistant principals) as well as all teachers in grades Prekindergarten (PreK) through 8 ($n = 888$). I chose to include elementary classroom teachers in those grades (i.e., self-contained general education, special education, and special areas [i.e., art, music, and physical education]) so that teachers classified as either Group A or Group B would both be represented. For the purpose of evaluation, elementary teachers for whom student achievement data are available to calculate teacher-level value-added scores are classified as Group A. Teachers for whom only grade- or school-level data scores can be used to assess their purported effectiveness as achievement data for their individual students are not available are classified as Group B.

Although high school administrators and teachers as well as other certified elementary staff (e.g., counselors, librarians, mathematics and reading interventionists,

instructional growth teachers [coaches]) also use the Danielson FFT and were included in the district-level evaluation, I chose not to include those groups in this study in an effort to narrow the focus of the research questions and to compare the perceptions of stakeholder groups in parallel positions at the elementary level (i.e., school principals and assistant principals as administrators and classroom teachers).

After obtaining approval from the Arizona State University (ASU) Institutional Review Board (IRB) and the district superintendent to conduct the study, I utilized district employee records to identify all elementary school administrators and teachers actively employed by the district as of January 13, 2014 (see Appendices D and E). Available records included employment data (i.e., primary job title, primary worksite, employment status [e.g., active, inactive, etc.], hire date, employee category [e.g., administrative, certified, etc.], and grade assignment if applicable) as well as demographic data (i.e., gender, race/ethnicity, total years of experience, and years of experience in the district). All personally identifiable information including employee names and district identification numbers were removed from the data file to protect confidentiality. Each employee in the population was assigned a unique identification number for the purposes of this study.

Phase 1: School Administrator and Teacher Interviews

Participant sampling. In Phase 1, I conducted interviews with randomly selected school administrators and teachers to inform the development of the surveys.

Specifically, I sought to assess the applicability of the standards of effectiveness as a conceptual framework, make defensible decisions with regards to the emphasis placed on each standard of effectiveness, ensure that appropriate terminology was used, etc. I

believe that adapting and adjusting the standards for the surveys when appropriate based on interview data collection and analysis served to mitigate, at least in part, the risk of restricting participants' abilities to express their own beliefs and opinions on the subject matter.

Acknowledging the potentially biasing impact of previous interactions with school administrators on the selection process, I deliberately chose to randomly sample participants. To complete the sampling process, I alphabetized the two lists of administrators (principals and assistant principals) and used a random number generator to rank order those in each group. I sent a letter via district email to the first five administrators on each list and invited them to participate in an interview (see Appendix F). I included an informed consent form for review and sent one reminder letter if I did not receive a response (see Appendices G and H). I recognized the ethical risk of an imbalanced power relationship with administrators who might feel compelled to participate (Hammack, 1997). In an effort to conduct the study in accordance with professional standards, the letter explicitly stated that participation was voluntary and confidential (i.e., in no way an employment expectation) as well as outlined the benefits (e.g., assuming an active role in organizational improvement) and (lack of) foreseeable risks to participation. In the event that an administrator did not respond or declined to participate ($n = 4$), I contacted the next administrator on the list. In total, I conducted ten interviews with administrators across seven schools (five principals and five assistant principals, respectively).

I repeated the same sampling process for elementary teachers, initially randomly selecting three participants from each of the following: Group A (grades 3-6), Group B

(grades PreK-2), and Group B (special areas). Each selected teacher received the same letter and informed consent form via email. Again, the letter advised teachers that participation was voluntary and confidential. To protect confidentiality, school administrators were not informed whether or which individual teachers on their campus had received (or declined) an invitation. In the event that a teacher did not respond or declined to participate ($n = 10$), I contacted the next teacher on the list for that group. In total, I conducted interviews with nine teachers across eight schools (three in each of the aforementioned groups).

Although I randomly sampled administrator and teacher participants, I acknowledged that their decision to participate was voluntary, and in turn recognized the potential for response bias in my sample (Daniel, 2012). As I did not plan to use the data to generalize to the population of school administrators or teachers but rather for the purposes of survey development, it was not imperative that the interview participants represented the population. Rather, I collected data until reaching a point of saturation (Strauss & Corbin, 1998).

Interview protocol. I developed separate but parallel semi-structured interview protocols for school administrators and teachers in an effort to evaluate the utility of the standards of effectiveness defined in my conceptual framework as a means of better understanding their perceptions and experiences in the local context (Guba & Lincoln, 1994) (see Appendices I and J). Throughout this process, I acknowledged the potential disjunction of the etic (outsider) theory I had imposed and the emic (insider) views of the participants and valued the interview experience as an opportunity to gain “rich insight”

into the “meaning and purposes attached by the [participants as] human actors to their activities” (Guba & Lincoln, 1994, p. 106).

The protocols were comprised of four main sections with series of open-ended questions not only intended to help me better evaluate the utility of the standards of effectiveness but also to provide opportunities for the participants to elaborate (see Figure 3). The protocols included 20 open-ended questions across the four sections:

- 1) Purpose of the Teacher Evaluation System (i.e., organizational – system design and implementation, and individual – evaluation process)
- 2) Measuring Teacher Quality (i.e., organizational – content adequacy of the Danielson FFT and value-added model as measures of teacher quality, and individual – professional practice and value-added scores)
- 3) Impact on Professional Practice (i.e., organizational – teacher hiring and retention and community perceptions, and individual – as a person and professional)
- 4) Improving Implementation (i.e., organizational – system as a whole, and individual – evaluation process).

Standard	Interview Question(s) (with protocol section)
<i>Effectiveness</i> (Purpose)	<ul style="list-style-type: none"> • What is the <i>purpose</i> of the teacher evaluation system? (S1) • Do you believe teachers share a <i>common understanding</i> of the purpose of the system? Why or why not? (S1)
<i>Fidelity of Implementation</i>	<ul style="list-style-type: none"> • To what extent have the processes for designing and implementing the evaluation system at the district-level been <i>transparent</i>? (S1) • How much <i>input</i> do you feel teachers have had in the process? (S1) • Do you believe the evaluation steps have been <i>clearly defined</i>? (S1) • What part of the evaluation process do you <i>value</i> most? Why? (S1) • Do you feel school administrators are well <i>prepared</i> to evaluate teachers? Why or why not? (S1) • In what ways, if any, can teacher evaluation be <i>improved</i>? (S4) • What <i>additional training</i>, if any, would be helpful for you in terms of the Danielson FFT rubric or teacher evaluation process? (S4)
<i>Popularity</i>	<ul style="list-style-type: none"> • Do you believe the <i>Danielson FFT</i> measures the most important aspects of teacher quality? Are there any domains or components you think are missing? (S2) • Do you believe the <i>value-added model</i> is a good measure of teacher quality? Why or why not? (S2) • If teachers' professional practice scores and value-added scores are <i>not aligned</i>, in which would you place more confidence? Why? (S2) • In general, do you believe that your final classification/label (based on the pilot year data) reflects your <i>level of effectiveness</i>? (S2) • Do you believe there is <i>consistency</i> among evaluators across the district? Why or why not? (S2) • Is the teacher evaluation process applied <i>fairly</i> to all teachers? (S2)
<i>Adaptiveness</i>	<ul style="list-style-type: none"> • How has participation in the teacher evaluation process impacted <i>you</i> professionally and personally? (S3) • Has it changed <i>your</i> professional practice? If so, in what ways? (S3) • Has the evaluation process impacted the professional practice of (other) <i>teachers</i> at your school thus far? If so, in what ways? (S3)
<i>Longevity</i>	<ul style="list-style-type: none"> • What impact, if any, do you think teachers' final classifications or labels will have on <i>teacher hiring and retention</i> at your school? (S3) • What impact, if any, do you think teachers' final classifications/labels will have on the <i>perceptions</i> of parents, students, and others in the community? (S3)

Figure 3. Alignment of Interview Protocols with Standards of Effectiveness (Cuban, 1998).

The protocols included several types of descriptive questions based on the standards of effectiveness and were intended to encourage the participants to describe their perceptions of the teacher evaluation system as a cultural scene (in terms of its purpose, implementation processes, impacts, etc.) (Spradley, 1979). For example, I asked both administrators and teachers grand tour questions (e.g., what is the purpose of the teacher evaluation system?) to better understand their perceptions in general (Spradley, 1979). Additional related mini-tour questions (e.g., to what extent have the processes for designing and implementing the evaluation system at the district-level been transparent?) and experience questions (e.g., what part of the teacher evaluation process do you value most and why?) provided interviewees with opportunities to share rich, more detailed descriptions of their personal beliefs and experiences (Spradley, 1979).

Data collection. In an effort to establish rapport with each interviewee, I offered to conduct the interview where he or she would be most comfortable (e.g., school site, district office, etc.) (Spradley, 1979). I explained the process and answered any questions before obtaining consent and beginning the interview. Each interview lasted approximately 45 minutes to an hour and was audio recorded with the participant's written permission. During the interview, I encouraged each participant to elaborate on his or her responses if so inclined and offered an opportunity to him or her at the end of the interview to address any additional aspects of the teacher evaluation system he or she felt were important. The semi-structured nature of the interview protocols allowed me to better understand how these participants measure the success of the teacher evaluation system.

Data management. Given the large amount of qualitative data collected through the interview process, I developed a strategy for organizing, managing, and securely storing files (Huberman & Miles, 1994). In order to ensure that I would have adequate time available for interview data analysis and survey development, I utilized a professional transcription service for interview audio files. In total, audio files from 19 interviews yielded 308 single-spaced pages of data transcribed verbatim. I organized audio files and transcripts using unique study identification numbers assigned to respective participants and read each transcript in its entirety to remove personally identifiable information (i.e., references to individual or school names). I utilized the web-based analytical software program Dedoose (2014) as my primary data management and storage system, uploading each transcript as a Microsoft Word document and attaching associated descriptors (i.e., position [administrator or teacher], gender, race/ethnicity, years of experience, and membership in Group A or B if applicable) to the data before beginning analyses. All qualitative data were stored in password protected folders and systems throughout this process.

Data analysis. Through collection and analysis of qualitative interview data, I sought to explore the perceptions of elementary administrators and teachers across the district prior to survey development and subsequent statistical analysis in the next phase. Although inductive designs are appropriate to explore unfamiliar, often complex cases, Huberman and Miles (1994) suggested that a researcher who is well acquainted with the study context; has a well-defined conceptual framework; and uses multiple, comparable cases to explain or confirm the applicability of concepts also has a “tight” qualitative design (p. 431). In this sense, I anticipated the need for data reduction by my “choice of

conceptual framework, of research questions, of samples, and of the ‘case’ definition itself, and of instrumentation” (Huberman & Miles, 1994, p. 430). As these choices inherently focused my work, I was not able to (nor should I have) simply analyze(d) interview data without recognition of the conceptual framework.

The conceptual framework provided key constructs (i.e., standards of effectiveness) for initial exploration via the interviews; however, the iterative task of analysis still required a measure of creativity (Huberman & Miles, 1994). Although not entirely atheoretical in nature as to warrant the use of grounded theory as an approach (Glaser & Strauss, 1967; Strauss & Corbin, 1995), I used the constant comparative method to generate an integrated, plausible analysis of the qualitative data (Erickson, 1986), all the while recognizing both convergent and divergent findings (Greene, 2007). After reviewing the entire corpus of interview data three times to begin discovering key linkages between and among the multiple data sources, I began to draw and substantiate with evidence my working assertions (Erickson, 1986; Smith, 1997). More specifically, by coding the text within Dedoose (2014) and identifying instances or basic units of analysis (Erickson, 1986), I determined the frequency with which codes appeared using a code calculation spreadsheet (Miles & Huberman, 1994) (see Table 1) and collapsed the code clusters into a series of major and minor themes.

Table 1

Qualitative Interview Analysis: Frequency of Themes by Position

Theme	Position		<i>n</i>
	Administrator	Teacher	
Purpose	47 (9.5%)	34 (8.3%)	81 (9.0%)
Evaluation System Structure	124 (25.1%)	92 (22.5%)	216 (23.9%)
Content Adequacy	3 (0.6%)	6 (1.5%)	9 (1.0%)
Evaluation Process	83 (16.8%)	93 (22.7%)	176 (19.5%)
Impact	103 (20.9%)	70 (17.1%)	173 (19.2%)
Evaluation System Improvements	75 (15.2%)	63 (15.4%)	138 (15.3%)
Agency	59 (11.9%)	51 (12.5%)	110 (12.2%)
Grand Total	494 (100.0%)	409 (100.0%)	903 (100.0%)

Note. Responses are presented as raw numbers with respective valid proportions of the total in parentheses.

Based on the frequency with which I applied codes to interview data and the themes I generated, I developed preliminary assertions to inform survey development (Strauss & Corbin, 1998). The sequential phases of data collection and analyses allowed me to adapt the conceptual framework when developing the surveys.

Phase 2: School Administrator and Teacher Surveys

Conducting a census. In Phase 2, I administered two separate but parallel online surveys to all elementary school administrators and teachers in the district. When initially designing the study, I considered the benefits and limitations of sampling versus census techniques for survey administration. Czaja and Blair (2005) explained that a survey is “based on the desire to collect information (usually by questionnaire) from a sample of respondents from a well-defined population” (p. 3). A probability sample allows a researcher to use information obtained from a survey instrument to generalize to or make inferences about the population of interest (Czaja & Blair, 2005). However, I hesitated to

randomly sample school administrators and teachers for two primary reasons. First, the small n count of elementary school administrators in the district ($n = 38$) would have required a sample nearly the same size as the population ($n = 35$) to use inferential statistics (Creative Research Systems, n.d.; Daniel, 2012). Second, the district placed greater value on the survey as an opportunity for all participants to describe their experiences, voice opinions, and share concerns. From an ethical standpoint, a census provided every school administrator and teacher in the district equal, albeit indirect access to the decision-making process (Daniel, 2012).

A census can be useful to achieve a representative sample, especially if a high participant nonresponse rate is a concern (Daniel, 2012). I had little evidence from which to predict the nonresponse rates for the surveys (e.g., nonresponse rates to previous district-wide surveys were not readily available), but I reasoned that the response rates might be higher if the surveys were conducted district-wide. I discuss the representativeness of participants in greater detail as part of the data collection section.

Survey instruments. I began the teacher survey development process by organizing the seven themes generated through interview data analysis (i.e., purpose, evaluation system structure, content adequacy, evaluation processes, impact, evaluation system improvements, and agency). I slightly redefined the standards of effectiveness in the conceptual framework based on my preliminary assertions. (I discuss the alignment of the final survey instrument and adapted standards in the next section.) As part of the survey development process, I wrote an initial draft in Microsoft Word and made significant subsequent revisions before creating an online version. As part of the revision process, I engaged in dialogue with district administrators and staff, soliciting their

feedback on various aspects of the survey structure, organization, and language. After I drafted a comprehensive teacher survey, I created a similar survey for administrators. The surveys were parallel in terms of structure although wording varied slightly to reflect positions and evaluator/evaluatee roles.

Although I recognized the value of field testing the instruments (Creswell, 2003), I could not collect pilot survey data directly from potential participants without subsequently removing them from the survey distribution list. In an effort to solicit additional feedback before making final revisions, I relied upon the critical review of an expert panel comprised of several district administrators and staff that had previously held positions as elementary school administrators or teachers (Czaja & Blair, 2005). In total, the panel included two former elementary principals and eight former elementary classroom teachers (i.e., including those who taught in self-contained general education, special education, Prekindergarten, and Structured English Immersion [for students with limited English proficiency] classrooms). They provided written feedback on the email letter to participants, survey structure, question sequence, language/word choice, etc. Dr. Audrey Amrein-Beardsley, my dissertation chair, also reviewed and provided specific written feedback on these aspects of the survey. Based on their recommendations, I made revisions in the surveys to improve question clarity, define terminology, etc. where necessary (Creswell, 2003).

The final administrator and teacher surveys included series of closed- and open-ended questions in six major sections and two additional sections with position-related and demographic questions organized as follows (see Appendices K and L):

- 1) Educator Position (two to four demographic items related to current position, assignment to a Title 1 school, grade level [teachers only], and membership in Group A or B [teachers only])
- 2) Purpose of Teacher Evaluation (two items related to the primary reasons for teacher evaluation in the district and in general)
- 3) Content Adequacy of Evaluation Measures (eight items related to the comprehensiveness of Danielson FFT, weighting of professional practice and student achievement components, non-test information and alternative achievement measures that should be considered, etc.)
- 4) Teacher Evaluation System Components (four items related to evaluation process fidelity, utility of evaluation activities, understanding of system components, sense of control over evaluation results, etc.)
- 5) Measuring Teacher Effectiveness (four items related to the fairness of the evaluation system, suggestions for improvement, professional development training needed, etc.)
- 6) Evaluation Implementation and Communication (four items related to the adequacy of district communication efforts, utility of resources provided, etc.)
- 7) Impact of Teacher Evaluation (four to six items related to the impact on professional practice [personally as an administrator or teacher, and on teachers in the school], student academic achievement, etc.)
- 8) Demographics (four demographic items for gender, race/ethnicity, total years of experience [teachers only], and years of experience in the district)

See the alignment of non-demographic questions to the adapted standards in Figure 4.

Standard	Survey Question(s) (with survey section)
<p><i>Effectiveness</i> (Purpose)</p>	<ul style="list-style-type: none"> • In your opinion, what <i>is</i> the primary reason (i.e., from a list provided) the district evaluates teachers' professional practice? (S2) • In your opinion, what <i>should be</i> the primary reason (i.e., from a list provided) for evaluating teachers' professional practices? (S2)
<p><i>Fidelity of Implementation</i></p>	<ul style="list-style-type: none"> • I feel <i>very comfortable</i> explaining to a non-educator how my/a teacher's ____ was calculated (i.e., professional practice score, value-added score, performance group assignment, overall effectiveness classification). (S4) • During this school year (2013-2014), which of these activities (i.e., from a list provided) were <i>conducted/completed</i> as part of your/teachers' evaluation(s)? (S4) • During this school year (2013-2014), how useful have each of the following evaluation activities (i.e., from a list provided) been in <i>helping you/teachers improve</i> your/their professional practice? (S4) • I believe that I/my <i>Administrator</i> have/has been <u>well trained</u> in the use of the Danielson rubrics to evaluate teachers. (S5) • I believe that I/my <i>Administrator</i> am/is able to evaluate teachers in an <u>objective and unbiased manner</u>. (S5) • The evaluation system would be significantly improved if: (S5) <ul style="list-style-type: none"> ○ Teachers were evaluated by <u>more than one observer</u>. ○ Teachers were evaluated by <u>an expert in their content area</u>. ○ Teachers were evaluated using <u>external evaluators</u>. ○ Teachers were evaluated (in part) by <u>peer-evaluators</u>. ○ <i>Administrators</i> received <u>more training</u> on the Danielson rubrics. ○ <i>Teachers</i> received <u>more training</u> on the Danielson rubrics. ○ The Danielson rubric criteria were <u>clarified or better defined</u>. • <u>Improvements</u>: In what ways, if any, could the teacher evaluation system or its implementation be improved? (S5) • <u>Professional Development/Information</u>: What additional professional development, training, or information (if any) related to the teacher evaluation system would be beneficial for you? (S5) • <u>Communication</u>: How well has the district informed/communicated with you regarding the development and implementation of the teacher evaluation system? (S6) • <u>Resources</u>: How helpful were the following resources (i.e., from a list provided) in improving your understanding of the purpose, design, and processes of the evaluation system? (S6)

	<ul style="list-style-type: none"> • <u>Classification Report</u>: Please indicate your level of agreement with the following statements about the Teacher Effectiveness Classification Report you/teachers received. (S6)
<p><i>Popularity</i></p>	<ul style="list-style-type: none"> • To what extent do the 22 components of the Danielson Framework for Teaching (FFT) incorporate <i>all/most</i> of the important characteristics of a good/effective teacher? (S3) • What, if any, important attributes/characteristics of good/effective teaching do you feel <i>should be added</i> to the evaluation system? (S3) • <u>Non-Test Information</u>: Should the district consider adding any of the following non-test information (i.e., from a list provided) to the evaluation criteria? (S3) • <u>Alternative Achievement Measures</u>: Should the district consider adding additional types of student achievement/learning measures (i.e., from a list provided) to the evaluation system? (S3) • When evaluating teachers, which of the following components provides the <i>best</i> indication of what it means to be good/effective (i.e., professional practice, student achievement, combination, or neither)? (S3) • A teacher’s overall evaluation score is currently computed as a combination of two primary factors: Danielson FFT rubric ratings (67%) and growth in student achievement (33%). In your opinion, how much weight <i>should</i> be given to each of these and to any additional components you believe should be represented? (S3) <ul style="list-style-type: none"> ○ Please explain what you meant by “other” and/or your rationale for assigning each of these weighting factors. (S3) • I believe that the <i>Overall Effectiveness Classification Label</i> I/teachers received for the 2012-2013 pilot year was an <u>accurate representation</u> of my/their professional performance. (S3) • I believe that the evaluation system <u>accurately captures</u> the impact teachers have on improving <i>student motivation, attitudes, and engagement in the learning environment</i>. (S5) • I believe that the evaluation system <u>adequately takes into account (adjusts for)</u> the influence of <i>student background characteristics</i> (i.e., demographics, prior achievement, program membership – special education, English language learner, gifted, eligible for free or reduced lunch) when determining my/teachers’ level of professional performance. (S5) • I believe that the evaluation system <u>fairly measures</u> the instructional/professional quality of teachers in <i>Group A</i>. (S5)

	<ul style="list-style-type: none"> • I believe that the evaluation system <u>fairly measures</u> the instructional/professional quality of teachers in <i>Group B</i>. (S5)
<i>Adaptiveness</i>	<ul style="list-style-type: none"> • I believe that I/teachers have <i>control over and can improve</i> my/their professional practice score/value-added score/overall effectiveness classification. (S4) • Overall, has the evaluation system had a positive or negative impact on <i>your</i> instructional/professional practices? (S7) • In what way(s) has the evaluation system impacted <i>your</i> professional practice? (S7) • Overall, has the evaluation system had a positive or negative impact on <i>student academic achievement and learning</i>? (S7)
<i>Longevity</i>	<ul style="list-style-type: none"> • If there is anything else that you would like to add about the <i>impact/consequences</i> of the evaluation system: please explain. (S7)

Figure 4. Alignment of Survey Items with Standards of Effectiveness (Cuban, 1998).

The administrator survey ($n = 27$ items total) had two more questions in the Impact on Teacher Evaluation section than the teacher survey ($n = 25$ items total). These additional questions related to the impact of teacher evaluation on each administrator’s own professional practice as well as that of teachers at his or her school.

Data collection. I created an online version of each survey in SurveyMonkey, a web-based software program with customizable survey design features including a web link (SurveyMonkey Inc., 2014). All 38 elementary school principals and 888 teachers received their survey link via email on May 5, 2014. Although I composed the contents of the email for each group, an administrator in the Human Resource Department assumed responsibility for sending the emails on my behalf with specific reference to this study (see Appendices M and N). This decision was made to ensure clear communication to potential participants, specifically that the district had authorized the use of district email accounts for the purposes of data collection. The emails explained the purpose of the study and need for research in this area, explicitly stated that participation was

voluntary and confidential, and provided instructions for survey submission (Plummer, 2001). Participants were permitted to complete the survey during non-instructional school hours if they chose to do so. All original recipients received three additional emails with the survey link (one per week) to remind them to participate (Czaja & Blair, 2005) (see Appendix O). The survey instrument remained open one week past the end of the school year. No additional responses were accepted after May 31, 2014.

Response rates. I began the data analysis process by determining the total response rate for each survey and comparing the participants to their respective populations for representativeness. In total, 76.3% of elementary school administrators ($n = 29/38$) and 76.0% of elementary teachers ($n = 675/888$) responded. As part of the district-level data collection process, high school administrators and teachers had received the survey link as well (i.e., via email based on current employment records); however, these respondents are not included in the response rates for this study above. I was able to determine response rates for elementary participants using a merged data file containing district employment and survey demographic data (with a dummy variable for participant/population status and a unique identification number for each record to protect confidentiality). To identify only elementary school administrators, I eliminated respondents ($n = 14$) who *did not* identify themselves as an elementary principal or assistant principal. Similarly for teachers, I eliminated respondents ($n = 376$) who *did not* both identify themselves as a general/special education or special area teacher (e.g., instructional support and related services staff) and indicate that they taught students in one or more elementary grades (PreK-8). Based on the remaining records, I determined that the sample size of teachers needed to support generalization was achieved. The

margins of error at the 95% confidence level for the teacher sample were +/- 1.85 (Creative Research Systems, n.d.).

Sample representativeness. Although relatively high response rates were achieved for both surveys, one could argue that some potential participants were uncomfortable answering questions related to district employee evaluation policies and practices despite assurances of anonymity (response bias). Although response rates exceeding 70.0% support claims of sample representativeness (Nunnally, 1978), I examined employment and demographic data more closely, calculating proportions within participant and population groups based on several characteristics (i.e., position, school eligibility for Title 1 funding, Group A or B membership [teachers only], gender, race/ethnicity, and years of experience) (Creswell, 2003). I tested equivalence using chi-square tests for homogeneity (Herringa, West, & Berlund, 2010). School administrator and teacher participants were not statistically significantly different from their respective populations in terms of position, school eligibility for Title 1 funding, Group A or B membership [teachers only], gender, race/ethnicity, and years of experience in the district. However, teacher participants and the population were statistically significantly different in terms of total years of experience (which includes experience outside the district), $\chi^2 = (5, N = 1450) = 24.09, p = .00$. Teachers with relatively fewer total years of experience (i.e., less than 13 years) responded at slightly lower rates than would have been expected based on their representation in the population. This, as discussed in the findings section, may have implications for the way in which findings are understood and should be kept in mind. Data for total years of administrator experience (which would

have included experience outside the district) were not available. Respondent demographic characteristics are discussed in greater detail in Chapter 5.

Data management. I applied a similar strategy for organizing, managing, and securely storing survey data as was described previously (Huberman & Miles, 1994). I exported quantitative and qualitative data from each survey from SurveyMonkey as Microsoft Excel spreadsheets (SurveyMonkey Inc., 2014). For the purposes of analysis, I organized and maintained data from each survey in a separate file. As all responses were anonymous, I assigned each record in the data files (i.e., all responses from one participant) a unique study identification number. Before beginning analyses, quantitative data (responses to demographic, close-ended, and Likert-type questions) were imported into IBM SPSS statistical software. I again used Dedoose (2014) to organize qualitative data by question, labeling each response with a position descriptor provided by the participant in the survey (i.e., school administrator or teacher). Before beginning systematic analyses, I read each response once in its entirety to remove personally identifiable information (i.e., references to individual or school names). All data were stored in password protected folders and systems throughout this process.

Data analysis. Through collection and analysis of qualitative interview data, I had developed surveys to better understand the perceptions of school administrators and teachers. The conceptual framework and associated research questions also guided analysis activities for both quantitative and qualitative survey data.

Quantitative data. I analyzed quantitative data for each survey separately and sequentially, beginning with the school administrator survey. I first calculated for each question in order the proportion of participants who chose each response option as well as

means and standard deviations for questions with Likert scales. I then repeated the same process for the quantitative teacher survey data. I also identified specific salient questions in the teacher survey for additional analysis and calculated proportions of Group A and Group B teachers who selected each response option. For ease of analysis, I created a table for each question in Microsoft Excel to organize and compare the proportions of selected response options by group. This data display technique helped me avoid “data overload” (Miles & Huberman, 1994, p. 56). I provide and discuss quantitative survey results for each question in Chapter 5.

As I began to visually examine aggregated data for each group by question, I discerned some variation in response option patterns by proportions of school administrators and teachers. However, I did not observe differences between the responses of teachers in Groups A and B similar to those observed between school administrators and teachers. Again, based on my initial review of the data, proportional responses of teachers in Groups A and B were closer than I would have expected. As I had aligned each question to a standard in the conceptual framework, I sought to better understand the possible variation in perceptions between school administrators and teachers through the qualitative survey data.

Qualitative data. When importing qualitative survey data into Dedoose (2014), I attached a position descriptor (i.e., school administrator or teacher) and question number to each response. This information allowed me to determine the number of written responses by group and question. In total, school administrators and teachers collectively submitted 837 written responses (i.e., 451 responses across four open-ended items and an

additional 386 explanations for selecting an “other” response option across eight closed-ended items).

Given the large volume of qualitative data collected, I utilized a structural coding process, described by MacQueen, McLelland, Kay, and Milstein (1998) as a means to identify “all of the text associated with a particular elicitation or research question” (p. 33). MacQueen et al. (1998) noted the utility of using structural coding with large databases because the index codes based on the research question or instrument structure can be applied to large segments of text. This process allows the researcher(s) to subsequently analyze within or across segments based on the index code (MacQueen et al., 1998).

Although MacQueen et al. (1998) indicated that it is not necessary to create index codes for open-ended survey data given the structured nature of survey questions, I created an index code for each of the seven broad themes that had been generated during interview analysis and used to develop the surveys as a starting point (i.e., purpose, evaluation system structure, content adequacy of measures, evaluation process, impact, evaluation system improvements, and agency) (see Appendix P). However, it is important to note that MacQueen et al. (1998) cautioned against the use of “professional jargon,” explaining that these etic codes tend to reflect the perceptions of the researcher(s) rather than the voice of the participants (p. 33). I recognized this to be the case in my own coding structure. For example, the code “Content Adequacy” in reference to the adequacy of various evaluation system components (i.e., the Danielson FFT) in measuring teacher effectiveness would not likely be used by most participants to appraise that system component. Although, in most cases, I applied one index code to each complete

participant response to a single question, I attempted to mediate my concern with etic codes by applying two or more index codes if helpful in capturing the essence of the response.

Additionally, and as suggested, I relied upon respondents' own language to construct emic codes for more in-depth analysis (MacQueen et al., 1998, p. 33). For example, I used the index code "Evaluation Process" to identify references to processes and procedures involved in conducting or participating in an evaluation and then applied the code (i.e., category) "Fairness" to responses that referenced any aspect of fidelity to or application of processes and procedures as inconsistent, biased, unfair, etc. By also adding another layer of codes(s) to those segments (i.e., already coded as "Evaluation Process" and "Fairness"), I was able to further identify specific references to fairness issues (e.g., "Adequate Time," "Evaluator Bias").

By using software in the manner suggested by MacQueen et al. (1998), I created a hierarchical coding structure during this process. The resulting database of coded responses was organized such that each response (row) had columns for the participant position and question identifiers as well as every code. The cell for each row and code column had a value of "0" or "1" to indicate whether that code had been assigned to that response. This feature allowed me to tabulate the frequencies of codes by hierarchical level and specific codes across questions by group. After reviewing the quantitative and qualitative survey data several times to discover key linkages, I revised my preliminary assertions (Erickson, 1986). As part of this process, I engaged in additional research activities intended to identify, acknowledge, and, if possible, address threats to the validity, reliability, and generalizability of study findings.

Validity

Validity, reliability, and generalizability merit careful consideration in mixed methods research. Using a pragmatic approach, I designed this study to be “methodologically appropriate” for the research questions and practical considerations associated with conducting social science research in this particular setting (Patton, 1990, p.72). Accordingly, I focused on quality judgment criteria appropriate for the methods used in this study and assessed context-specific threats to validity, reliability, and generalizability.

Recognizing that definitions of validity are often paradigmatic, Creswell and Miller (2000) defined validity as “how accurately the account represents participants’ realities of the social phenomena and is credible to them” (p. 124; see also Schwant, 1997). Accordingly, they suggested three particular lenses for examining validity, namely that of the researcher, participants, and peer reviewers (Creswell & Miller, 2000). Respective lens are paradigmatic as well and as such favor specific validity procedures (Creswell & Miller, 2000). For the purposes of this study, I employed procedures suited for each of the three lenses, specifically emphasizing those of the researcher and participants.

Researcher lens. Searching for convergence among multiple, different data sources (i.e., interview and survey data from both school administrators and teachers) as a means of validation, I used both data (i.e., participants) and methods (i.e., instruments) triangulation to justify the coherence of themes (Denzin, 1978). Throughout the study, I maintained a detailed log of research activities (e.g., including the date, type of activity, data source/participants, instrument used, data file structure, software/analytic tool). I

reviewed the log when triangulating data from multiple sources for ease of analysis and to avoid interpretation errors. For example, I could have attributed an instance from qualitative interview data in Dedoose (2014) to the wrong participant group (school administrator or teacher) by misinterpreting a participant descriptor field. Carefully documenting research activities allowed me to systematically read and reread the entire corpus of interview and survey data while seeking confirming and disconfirming evidence (Miles & Huberman, 1994).

As part of the triangulation process, I sought to identify patterns within and across the cases (i.e., school administrators and teachers as groups) from multiple sources without trying to justify generalization from one group/source to another (Erickson, 1986). Miles and Huberman (1984) suggested that researchers are more likely to identify evidence that reinforces their original beliefs, a phenomenon I recognized in my own triangulation process. In an effort to overcome this challenge, I purposefully examined data for negative or extreme cases. To ensure representativeness of the illustrative instances to be presented in Chapter 5, I identified and organized both exemplary (confirming) and negative/extreme (disconfirming) instances/quotations from participants in tabular format for each theme (Kane, 2001, 2013).

Based on the tables of confirming and disconfirming evidence, I supported each preliminary assertion using one or more of the following: 1) a synopsis of patterns [“general description”]; 2) basic instances of analysis from the interviews and/or surveys [“particular description”]; and 3) my own interpretive commentary. This format served a two-fold purpose: to define for readers (in this case, district administration) what I meant by each assertion and to provide supporting evidence (Erickson, 1986, p. 149). As

suggested by Erickson (1986), I utilized general and particular description in tandem to provide evidence supporting my assertions both in terms of “breadth of evidence” and “relative frequency of occurrence of a given phenomenon” (p. 149). For example, I asserted that school administrators and teachers held different beliefs about the primary purpose of the district’s evaluation system. As evidence for this assertion, I prepared descriptive statistics in tabular format for the two survey questions specifically related to the purpose of evaluation and direct quotations that were exemplary of responses from school administrators and teachers (identified by group).

I considered Erickson’s (1986) argument that even assertions that seem “believable on intuitive grounds” are open to criticism if not supported by systematic analyses of data (p. 149). I sought to provide evidence that the “patterns of generalization [I had identified within and across groups] within the data set” were plausible by “account[ing] for patterns found across both frequent and rare events” (i.e., common themes based on code frequency as well as basic instances of analysis representing divergent themes) (Erickson, 1986, p. 149). The triangulation and reporting processes were critical steps in establishing the validity or credibility of study findings.

Participant lens. In order to engage in dialogue with participants as a means of establishing credibility (akin to internal validity) (Lincoln & Guba, 1985), I conducted a focus group with teachers from across the district on August 28, 2014. According to Lincoln and Guba (1985), these member checks are “the most crucial technique for establishing credibility,” and if a researcher “purport[s] that his or her reconstructions are recognizable to audience members as adequate representations of their own (and multiple) realities...[then they must] be given the opportunity to react” to the findings (p.

314). Erickson (1986) also suggested that the most salient concerns of practitioners include “deciding whether or not the situation described in the report has any bearing on the situation of their own practice” (p. 153). As practitioners, this focus group represented another audience with its own interest in study results and findings (Erickson, 1986).

In an effort to solicit practitioners’ responses to my preliminary assertions, I relied on a list of teachers (one from each school) who had participated in a teacher evaluation focus group in fall 2013 to identify potential participants for this group. I purposively chose to use this list of teachers from respective grade bands (i.e., PreK-2, 3-6, middle level Group A [grades 7-8 mathematics and English language arts], and middle level Group B [grades 7-8 science and/or social studies and special areas]) as they had already demonstrated their willingness to represent their peers and each had had an opportunity to participate in the survey in spring 2014. It is important to note that four additional high school teachers were included in the original list and subsequently invited to attend the focus group as study findings were relevant to the larger district-level evaluation. I composed an invitation explaining the purpose for the group and limits of the information desired. I knew that communicating a clearly defined purpose to participants and narrowing their focus to essential topics was critical to conducting an effective focus group (Jarrell, 2000). Again, the same administrator from the Human Resource Department assumed responsibility for sending the email on my behalf.

In total, slightly more than half of invited elementary teachers ($n = 11/19$, 57.9%) attended the focus group (i.e., PreK-2, $n = 2$; grades 3-6, $n = 7$; middle level Group A [grades 7-8 mathematics and/or English language arts], $n = 1$; and middle level Group B [grades 7-8 science and/or social studies and special areas], $n = 1$). The number of focus

group participants was slightly larger than the recommended four to six members (Tynan & Drayton, 1988; Vaughn, Schumm, & Sinagub, 1996). Holding the focus group after school likely increased the number of teachers able to attend (Vaughn et al., 1996). The focus group was structured such that teachers were asked to provide three levels of feedback: 1) judgment of overall credibility, 2) statements about major concerns or issues, and 3) and statements about factual or interpretive errors (Lincoln & Guba, 1985).

During the focus group, I presented preliminary study findings and asked each teacher to independently complete an online form with specific questions intended to capture his or her reaction to my preliminary assertions. In the form, each teacher identified his or her grade band and answered the first four questions independently before engaging in group discussion. After the discussion, each teacher answered one additional question. The focus group questions were as follows:

- 1) Now that you have seen an overview of study findings, please share your initial thoughts and feedback.
- 2) Do the study findings resonate with your personal experience? Please explain.
- 3) Do the study findings make sense based on any conversations you have had with other teachers at your school? Please explain.
- 4) What aspects of the evaluation system and/or process should be the focus for improvement?
- 5) After you have participated in the group discussion, please provide any additional thoughts or feedback.

Given the semi-structured format of the focus group protocol and equal opportunity for written feedback, the larger group size did not seem to negatively impact

or limit teachers' ability to participate. After conducting the focus group, I exported qualitative data from the online form to an Excel spreadsheet. It is important to note that two high school teachers attended the focus group as well; however, their written feedback was excluded from analyses for the purposes of this study. I applied the same coding scheme used previously to analyze the interviews and surveys (Miles & Huberman, 1994). By using the same coding scheme, I was better able to authentically compare multiple data sources (i.e., interview, survey, and focus group data) and continuously seek disconfirming evidence (Erickson, 1986; Miles & Huberman, 1984, 1994; Smith, 1997).

Based on their written responses, most focus group participants reportedly found that the study findings resonated with their personal experiences, specifically with regards to differences in the perceptions of school administrators and teachers. One teacher wrote, "I find the results interesting but not entirely surprising. I think there is a schism between the viewpoint of the person delivering the evaluation and the person on the receiving end of the evaluation." Another teacher expressed his or her agreement but also questioned the willingness of respondents to answer survey questions honestly:

I am glad to see that my thoughts and feelings about the evaluation system have been validated, but I wonder how many teachers and administrators, based on the results, were afraid to be honest when completing the survey. It was interesting that administrators had in some areas a completely different perspective of the impact on the evaluation system than the teachers in the trenches.

Although most teachers were not surprised by study findings related to teacher perceptions, a few would not have anticipated perceptual differences with administrators.

One teacher explained, "...I felt the same way. I was surprised that principals felt certain ways [though]. You are always under the impression that they know the overall expectation." Another teacher concurred with the findings but expressed hope for improvement: "Overall, yes. There are teachers that do not feel they are truly represented in their summary of effectiveness. Maybe this year will be better as teachers become more familiar with the process."

Teachers also commented on the study findings in the context of their conversations with other teachers at their school. One teacher wrote about the inherent impact of teachers' evaluation results on their perceptions of the process:

Yes, the surveys make sense. Overall, if a teacher agreed with their [evaluation] results, they had positive comments about the process, and if they did not agree with their rating, they had negative responses. My peers were not surprised with their results and felt they wanted to make changes to improve their labels.

Another teacher reiterated the concerns of his or her peers as described in study findings: "Yes. I think that other teachers have the same concerns...there are many things that teachers feel they cannot control in this aspect, and in return there is frustration." Another teacher expressed similar sentiment, noting that "Yes, for the most part [study findings make sense]. The teachers at my school that [sic] I have talked to were not happy with their placements either or felt that it didn't really reflect their teaching practices." In addition, several focus group teachers anonymously described the negative experience of an individual peer during the evaluation process as evidence in support of study findings.

When asked what aspects of the evaluation system should be the focus for improvement, teachers expressed the importance of communicating with district and

school administrators. One teacher wrote that it is important that district administrators “continue to ask for teacher input and commentary. Teachers want to know that their voice is being heard, acknowledged, and considered.” Another teacher emphasized the need for continuous communication: “We need more dialogue and training. [A]nd more dialogue and more dialogue.” One teacher acknowledged the need for communication and understanding between administrators and teachers on a broader level, explaining:

I think there should be more dialogue between teachers and administrators about what characteristics of teaching we really value. What is it that makes us a better school? How can we work together to better the lives of our students? When these discussions take place and people can come to a common understanding and expectation, I think the rubric and achievement scores will take care of themselves.

Several other teachers expressed an overarching sense of hope that increased dialogue with school and district administration would improve not only the evaluation system and its processes but also, perhaps more importantly, acknowledge teachers as professionals, foster positive school environments, and increase student learning. Analyzing focus group participant responses in conjunction with interview and survey data enabled me to recognize a larger, conceptual web in which many of these themes overlapped (Erickson, 1986).

Peer reviewer lens. In an effort to examine evidence of validity with the assistance of peer reviewers, I also engaged in a debriefing process with researchers in the district who had been actively involved in the design and implementation of the teacher evaluation system (Lincoln & Guba, 1985). This review process occurred at

several junctures in study design and implementation, namely after each of the following: 1) establishing the research questions; 2) devising the sampling design; 3) developing the survey instruments; and 4) reviewing the preliminary assertions. Again, while I assumed primary responsibility for each of the aforementioned, continuous dialogue undoubtedly maximized the utility of the study as an improvement tool for district leadership. While the reviewers were not necessarily external, they provided invaluable support, critiqued study methods, and challenged my assumptions (Lincoln & Guba, 1985).

Reliability

In a positivist paradigm, validity does not exist without reliability; however, Lincoln and Guba (1985) suggested that the same argument for credibility in the absence of dependability (arguably parallel in naturalistic inquiry to reliability) is inadequate. Rather, dependability should be independently examined using techniques such as an audit (Guba & Lincoln, 1989; Lincoln & Guba, 1985). It is important to note that as an outcome of the evaluation, I prepared a written report outlining my preliminary assertions for district administration. In the report, I systematically described the research design, data collection methods, data analysis processes, preliminary assertions with supporting evidence, conclusions, and recommendations.

In an effort to audit both the research processes and the report as a study product (Lincoln & Guba, 1985), I held two extended meetings (approximately 3 hours each) in July 2014 with Education Services and Human Resource Department Directors. Each director was invited to examine the research log, raw (personally unidentifiable) qualitative data, aggregated survey results, and report in order to provide feedback. Their reactions to the data and preliminary assertions were particularly valuable as a means of

validating my own data processes, syntheses, and interpretations. After the focus group and district meetings, I reviewed the entire corpus of interview and survey data again for confirming and disconfirming evidence for each of my preliminary assertions (Erickson, 1986).

Finally, I prepared a presentation outlining my warranted assertions for district policymakers, namely members of the District Cabinet (i.e., the superintendent, assistant superintendents of education and support services, and executive directors of human resources and business services), and the District Governing Board (presented by a member of Cabinet at a public meeting on October 1, 2014). I believe that these culminating activities provided district leadership with an opportunity to reconceptualize teacher evaluation as a policy and identify options for system improvement by discussing the “unintended consequences of implementation, unanticipated barriers to it, and unrecognized reasons why it was successful” (or unsuccessful) in this setting (Erickson, 1986, p. 153). In consideration of the study findings, district administrators developed a plan for improving teacher evaluation processes and procedures. Specific components of the plan, including a formal study of inter-rater reliability and new teacher professional development trainings and materials focused on the Danielson FFT rubrics, are currently being implemented.

Generalizability

Although an adequate sample size is needed to generalize results to the population (as evidence of external validity) (Czaja & Blair, 2005), Guba and Lincoln (1989) denoted transferability as more appropriate for naturalistic inquiry and argued that it is “always relative and depends entirely on the degree to which salient conditions

overlap or match” with the burden of proof on the receiver rather than the inquirer (p. 241). In terms of mixed methods research, transferability is arguably applicable as well, again based on the study’s purpose and design. In this regard, the degree of transferability is evidenced by “thick description” (Geertz, 1973, p. 6) such that others can make judgments about whether the data and findings in a study can be applied to their own situation (Lincoln & Guba, 1989).

Stake (1978) suggested that the use of a process termed “naturalistic generalization” to apply the findings of a study to other similar situations is of equal value as a means of understanding, in this case, as per the perceptions and experiences of school administrators. Schofield (1993) further proposed targets for generalization that can, in this case, be applied to teacher evaluation processes and systems: *what is, what may be, and what could be*. I believe that this rich collection of qualitative and quantitative data captures respondents’ perceptions of what is, may be, and could be happening in the district both in terms of the conceptual framework and in their own words.

Study Limitations

I recognize that the conceptual framework and sequential nature of the design posed certain limitations. Strauss and Corbin (1998) described a theory as “more than a set of findings [as] it offers an explanation about phenomena” and cited the importance of applying concepts that explain phenomena in one context to another (p. 23). In this way, they argued that one can determine whether the concepts “might prove valuable...for explaining similar phenomena” in other contexts (Strauss & Corbin, 1998, p. 23). Because the predefined conceptual framework inherently shaped the research design for

this study, the use of surveys to better understand participant perceptions could arguably have restricted their ability to authentically articulate beliefs, opinions, and concerns.

Conducting interviews to better “understand the meaning and nature of [the] experiences” of school administrators and teachers in this context was essential as a means of informing survey development (Strauss & Corbin, 1998, p. 11). I relied upon interview data to inform the development of the surveys in terms of the emphasis placed on each standard of effectiveness, terminology used, etc. which served to mitigate, at least in part, the concern. Additionally, I engaged in formal and informal dialogue with district leadership regarding study findings and conclusions and conducted a teacher focus group at the end of the study to (dis)confirm my assertions and conclusions before making formal recommendations. I sought to carefully consider, acknowledge, and, if possible, address issues related to credibility (internal validity) and dependability (reliability) to enable others to determine the naturalistic generalizability of study findings to better understand the lived experiences of school administrators and teachers in other similar contexts.

CHAPTER 5

Results

In this study, I examined the perceptions of elementary school administrators and teachers regarding the purpose, implementation, effectiveness, and intended/unintended consequences of a state policy-directed, locally-designed and implemented teacher evaluation system. The primary research questions were:

- 1) What do stakeholders perceive as the purpose and goals of the locally-designed teacher evaluation system to be used in their district?
- 2) How do stakeholders describe the intended and actual implementation processes for the teacher evaluation system?
- 3) How do stakeholders measure the effectiveness of the teacher evaluation system based on their understandings of the purpose/goals as well as the intended and actual implementation processes?
- 4) To what extent do perceptions of the purpose/goals, descriptions of implementation, and measures of effectiveness vary across stakeholder groups?

In this chapter, I initially provide a detailed demographic description of survey participants before discussing study results in order of the research questions referenced above. I address Question #4 as part of the respective discussions for Questions #1-3 by presenting and contrasting the perceptions of administrators and teachers with regards to the system's purpose, implementation (i.e., intended and actual), measures of effectiveness (i.e., validity, reliability, and fairness), and intended/unintended consequences. It is important to note that the interview and survey instruments used in

this study were designed to capture school administrators' and teachers' lived experiences as those supposedly reforming and being reformed through the actual implementation phase of a larger, more complex policy cycle. In addition, I reflect upon the distinctiveness and nuances of their experiences and perceptions using their own voices to evidence results.

Demographic Description of Survey Respondents

School administrators. In total, 76.3% of elementary school administrators in the district ($n = 29/38$) responded to the survey. As mentioned previously, a response rate exceeding 70% supports claims of sample representativeness (Nunnally, 1978). Chi square tests of homogeneity also demonstrated that the school administrator sample and population were not statistically significantly different in terms of position, assignment to a school receiving Title I funding, gender, race/ethnicity, and administrator experience in the district. As a result, I primarily discuss the demographic characteristics of respondents. Population demographic characteristics are presented in Table 2.

Survey respondents identified themselves as either principals ($n = 15/29$, 51.7%) or assistant principals ($n = 14/29$, 48.3%) in nearly equal proportions. Slightly less than 40% of administrators ($n = 11/29$, 37.9%) indicated that they were assigned to a school receiving Title I funding. Of those who reported their gender and/or race/ethnicity, the vast majority identified themselves as female ($n = 23/26$, 88.5%) and/or Caucasian/White and not Hispanic/Latino ($n = 21/22$, 95.5%). More than half of respondents ($n = 16/26$, 61.5%) indicated that they had fewer than four years of experience as an administrator in the district. Respondents were also asked to report their total years of experience as an administrator (including years outside the district); however, I questioned the integrity

and interpretation of the data as the survey question did not specify whether they should include experience in other administrator positions (i.e., district-level or other non-school site positions). In addition, employment data for administrator-specific experience outside the district were not available for the population.

Table 2

Employment and Demographic Characteristics of School Administrators

Characteristic	Participants <i>n</i> = 29	Population <i>N</i> = 38
Position		
Principal	15 (51.7%)	20 (52.6%)
Assistant Principal	14 (48.3%)	18 (47.4%)
Total	29 (100.0%)	38 (100.0%)
Title 1 School		
Yes	11 (37.9%)	16 (42.1%)
No	18 (62.1%)	22 (57.9%)
Total	29 (100.0%)	38 (100.0%)
Gender		
Male	3 (11.5%)	9 (23.7%)
Female	23 (88.5%)	29 (76.3%)
Total	26 (100.0%)	38 (100.0%)
Race/Ethnicity		
White, Not Hispanic	21 (95.5%)	31 (81.6%)
Any Other Race/Ethnicity	1 (4.5%)	7 (18.4%)
Total	22 (100.0%)	38 (100.0%)
Years of Experience (in District)		
1-3	16 (61.5%)	17 (44.7%)
4-6	4 (15.4%)	10 (26.3%)
7-9	5 (19.2%)	9 (23.7%)
10-12	1 (3.8%)	2 (5.3%)
13-15		
16 or More		
Total	26 (100.0%)	38 (100.0%)

Note. Counts are presented as raw numbers with respective valid proportions of the total participants or population in parentheses.

Teachers. In total, 76.0% of elementary teachers ($n = 675/888$) responded to the survey. The high response rate and results of chi square tests of homogeneity also demonstrated that the teacher sample and population were not statistically significantly different in terms of position, assignment to a school receiving Title I funding, membership in Group A or B, gender, race/ethnicity, or teaching experience in the district. As a result, I primarily discuss the demographic characteristics of teacher respondents. Population demographic characteristics are presented in Tables 3 and 4.

The vast majority of respondents identified themselves as general education classroom teachers ($n = 597/675$, 88.4%), which included special area teachers (i.e., art, music, and physical education). More than four out of ten teachers ($n = 286/665$, 43.0%) indicated that they were assigned to a school receiving Title I funding, and 65.0% reportedly taught primary grades ($n = 211/674$, 31.3% in PreK-2 and $n = 236/674$, 35.0% in grades 3-6, respectively). The remaining one-third of teachers taught grades 7-8 ($n = 125/674$, 18.5%) or all/multiple grades ($n = 102/674$, 15.1%).

Although teachers were asked to identify whether they were members of Group A or Group B, I questioned the integrity of that data as well. Due to a technical error, some teachers were incorrectly designated as Group A in the 2012-2013 Teacher Effectiveness Classification Report they received shortly before survey administration. Although this reporting error did not affect value-added calculations or evaluation outcomes, some respondents likely indicated that they were members of Group A rather than Group B. To more closely estimate each group, all teachers who reported that they taught grades 3-6 were designated as Group A ($n = 236/675$, 35.0%). All other respondents were designated as Group B ($n = 439/675$, 65.0%). These proportions reflect others' estimates

of teachers in tested grades and subjects for whom student-level achievement data are available (Amrein-Beardsley, 2014; Harris, 2011).

Table 3

Employment Characteristics of Teachers

Characteristic	Participants <i>n</i> = 675	Population <i>N</i> = 888
Position (Classroom Teacher)		
General Education	597 (88.4%)	796 (89.6%)
Special Education	78 (11.6%)	92 (10.4%)
Total	675 (100.0%)	888 (100.0%)
Title 1 School		
Yes	286 (43.0%)	363 (40.9%)
No	379 (57.0%)	525 (59.1%)
Total	665 (100.0%)	888 (100.0%)
Grade^a		
PreK-2	211 (31.3%)	261 (32.4%)
3-6	236 (35.0%)	305 (37.8%)
7-8	125 (18.5%)	148 (18.4%)
All/Multiple (K-8)	102 (15.1%)	92 (11.4%)
Total	674 (100.0%)	806 (100.0%)
Group^b		
A	236 (35.0%)	351 (39.5%)
B	439 (65.0%)	537 (60.5%)
Total	675 (100.0%)	888 (100.0%)

Note. Counts are presented as raw numbers with respective valid proportions of the total participants or population in parentheses.

^aSome employee records did not specify grade assignment. Teachers assigned to work with multiple grades are likely underrepresented in the population N count.

^bDue to a technical error in the SY2012-2013 Teacher Effectiveness Classification Reports, some respondents likely incorrectly identified themselves as Group A teachers. To better estimate group membership, all participants who taught in grades 3-6 were designated as Group A. All other participants were designated as Group B. This calculation may underestimate the number of Group A teachers.

The vast majority of teachers who reported their gender identified themselves as female ($n = 458/549$, 83.4%). Of those who reported their race/ethnicity, 82.5%

identified themselves as Caucasian/White and not Hispanic/Latino ($n = 442/536$).

Slightly more than one-third ($n = 215/561$, 38.3%) identified themselves as *probationary*, meaning that they had fewer than four years of experience as a teacher in the district (including the 2013-2014 school year). The remaining teachers ($n = 346/561$, 61.7%) were considered *continuing* as they reported four or more years of in-district experience.

One-fifth of teachers ($n = 107/561$, 19.1%) reported ten or more years of in-district experience. It is important to note that many teachers came to the district with outside teaching experience. As such, the proportion of teachers who might be considered relatively new to the profession (i.e., with less than four total years of experience; $n = 94/562$, 16.7%) was much lower when outside experience was included. In fact, the majority of teachers ($n = 288/562$, 51.2%) reported more than ten total years of experience.

Table 4

Demographic Characteristics of Teachers

Characteristic	Participants	Population
	<i>n</i> = 675	<i>N</i> = 888
Gender		
Male	91 (16.6%)	147 (16.6%)
Female	458 (83.4%)	741 (83.4%)
Total	549 (100.0%)	888 (100.0%)
Race		
White, Not Hispanic	442 (82.5%)	752 (84.7%)
Any Other Race/Ethnicity	94 (17.5%)	136 (15.3%)
Total	536 (100.0%)	888 (100.0%)
Years of Experience (in district)		
1-3	215 (38.3%)	395 (44.5%)
4-6	107 (19.1%)	163 (18.4%)
7-9	132 (23.5%)	200 (22.5%)
10-12	64 (11.4%)	88 (9.9%)
13-15	32 (5.7%)	26 (2.9%)
16 or More	11 (2.0%)	16 (1.8%)
Total	561 (100.0%)	888 (100.0%)
Total Years of Experience		
1-3	94 (16.7%)	200 (22.5%)
4-6	84 (14.9%)	168 (18.9%)
7-9	96 (17.1%)	156 (17.6%)
10-12	87 (15.5%)	146 (16.4%)
13-15	83 (14.8%)	88 (9.9%)
16 or More	118 (21.0%)	130 (14.6%)
Total	562 (100.0%)	888 (100.0%)

Note. Counts are presented as raw numbers with respective valid proportions of the total participants or population in parentheses.

Research Question 1: Purpose of Teacher Evaluation

As discussed in Chapter 2, increasing public demand for school and teacher accountability has prompted both federal and state policy changes that require the development and implementation of teacher evaluation systems that rely, at least in part, on complex statistical models to estimate teacher effectiveness (Amrein-Beardsley, 2008, 2014; Harris, 2011; Papay, 2010). The locally-developed teacher evaluation system implemented in this district is comprised of both professional practice (67.0%) and value-added (33.0%) measures. To better understand the perceptions of school administrators and teachers, interviewees and survey respondents were asked to explain the purpose of the teacher evaluation system in the district. Analysis of interview data evidenced the divergent perceptions of school administrators and teachers regarding the system's purpose. Although not directly asked, interviewees also differentiated between the ideal purpose of evaluating teachers in general and the actual purpose of the system in the district. This distinction in the interview data informed the development of two parallel survey questions, specifically referencing the primary reason(s) for evaluation, ideally and in reality, as two separate domains.

In general, school administrator interviewees described the purpose of the evaluation system as a means to improve teachers' professional practice. One administrator explained, "I feel that it's to have teachers just improve in their craft and help give them direction for what we're going for in the district. Really guide them...but align what they're doing to what the district expects." Another administrator also described the system as a means to support teachers: "We want to be looking at the effectiveness of teachers and be able to give them meaningful feedback so that they can

take that feedback and put it into practice and improve their practice.” A third administrator concurred, specifically referencing the Danielson FFT as the common measure for identifying and supporting growth:

The purpose of the teacher evaluation system is to indicate to teachers what they are doing well and where there are opportunities to grow based on common themes, really, or common categories that are laid out in the Danielson rubric.

Although most administrators cited professional growth as the purpose of the evaluation system in the district, others juxtaposed the ideal purpose (i.e., improving professional practice) with what they believe to be the actual purpose (i.e., accountability).

Administrators who described a disjunction between the ideal and actual purposes of the system cited both the need to evaluate teachers to make employment decisions and to comply with federal and state policy mandates. Illustrating the multi-purpose nature of the system, one administrator explained that, “the primary purpose is to help teachers to be successful and become the best teachers they can be. It is also used on the negative side, if you have an issue where you need to move someone then it’s used for that too.” Another administrator contextualized the use of the evaluation system to make employment decisions, citing his or her own perspective as a parent:

I think we have teachers that people [administrators] haven’t done their jobs on. As an instructional leader, it’s my job to either help you get better or say, ‘Maybe this isn’t for you.’ Teaching isn’t for everyone. If you just want to come and work from 9:00 to 4:00, it’s probably not a really good match for you. When I look at my [teachers], like I said, who are ineffective, I wouldn’t put my own kids in that classroom. Then is it right for me to put anyone else’s kid in their classroom?

Another administrator noted that his or her colleagues use the system to encourage some teachers to reconsider their profession:

I think people [administrators] are using that to either coach or evaluate people into a different career, to be honest with you. We all have teachers that [sic] probably are struggling to get better and they're not really getting better. Some of them are early in their careers. Some of them have been around for a while, and they've been doing the same thing, and they're just kind of—what we don't want, I don't think, [which] is mediocre teachers.

While some administrators advocated for the use of evaluation results to make employment decisions, others noted the legislative impetus for the development and implementation of the system. One administrator explained his or her understanding as follows:

The evaluation system was a standardized approach to evaluate teachers and to evaluate teacher performance. It was a single approach that was basically one, required by the state, some form of it, and customized by our district so that we have an across-the-district comparison of teacher performance.

While some school administrators acknowledged the plurality of the evaluation system, teacher interviewees described the complexities of this phenomenon from their own perspective instead.

Most teachers interviewed expressed the concerns and frustrations of their peers regarding the purpose of the teacher evaluation system while reiterating their own belief in the system as a tool for professional growth. One teacher described the primary

purpose of the system as a means to facilitate professional growth through communication with his or her administrator. The teacher explained:

I think we [teachers] all have a common understanding, even though it's kind of changed. I've taught in three different states, and it's all pretty much been the same idea...this is the purpose, the self-reflecting and growing.... I feel that it's more of a conversation [with my administrator], which is how I think it should be.

However, the teacher then added, "I don't know that everyone uses it in the appropriate way."

Other teachers specifically described conflicting perceptions of the purpose among their peers. One teacher explained,

In my opinion, the purpose of the teacher evaluation system is to allow teachers to reflect on their own teaching through observations of others, and allow us the chance to improve ourselves, to make sure we are constantly doing a better job so that our students will be more successful.

However, he or she further explained:

[T]eachers feel that the evaluations are a tool that's going to be used against them in some fashion, either to—I don't know even how they could—but to decrease pay or to give them a less desirable position next year, or to punish them if they argued with an administrator somewhere.

Lastly, he or she added, "It's kind of frustrating to see that. The general attitude seems to be [that] evaluations are a negative thing rather than a way to grow." A third teacher responded similarly, noting "I know that for the purposes of the individual that's [sic] being evaluated it's to kind of tweak them or make them a better educator as well, where

they might need some support.” However, he or she described the employment consequences for teachers who “don’t make proficiency,” writing that “you’re basically at a fork in the road if you’re on an improvement plan. You either go this way to make yourself better and get off of it [the plan], or you go this way, and you end up bye-bye.” In conclusion, he or she acknowledged the larger purpose of the system in terms of educational policy, adding “I know that society, the community, wants us to have a better idea about what teachers are doing in the classroom and what the children are getting out of the learning. I get that.” As previously stated, school administrators and teachers through the interview process described their own complex and, to some extent, divergent perceptions regarding the ideal and actual purposes for evaluating teachers. These findings prompted the inclusion of two survey questions to address each domain separately.

Perhaps unsurprisingly, all administrator survey respondents ($n = 29/29$, 100%) indicated that the primary reason teachers *should* be evaluated is to improve their professional practice (see Table 5). The vast majority of teacher respondents ($n = 569/663$, 85.8%) agreed with this sentiment although it is worth noting that nearly one in ten teachers ($n = 54/663$, 9.7%) reported that the system should be used to hold them accountable. However, when asked to describe the district’s evaluation system in reality, slightly fewer administrators ($n = 23/29$, 79.3%) reportedly believe that the system *is* primarily intended to support professional growth. Reflecting disparaging perceptions, a substantially smaller proportion of teachers ($n = 233/662$, 35.2%) agreed that the district’s evaluation system is actually in place to improve their professional practice. In fact, more than four out of ten teachers ($n = 296/662$, 44.7%) indicated that the system

was implemented as an accountability mechanism. In addition, 13.4% of teachers ($n = 89/662$) cited compliance with state legislation as the primary purpose.

Table 5

Primary Reasons for Evaluating Teachers

Primary Reason	Administrators		Teachers	
	Ideally	This District	Ideally	This District
Improve professional practice	29 (100.0%)	23 (79.3%)	569 (85.8%)	233 (35.2%)
Hold teachers accountable		4 (13.8%)	64 (9.7%)	296 (44.7%)
Make employment decisions			7 (1.1%)	32 (4.8%)
Comply with state legislation		2 (6.9%)	2 (0.3%)	89 (13.4%)
Teachers should not be evaluated		-	9 (1.4%)	-
Other			12 (1.8%)	12 (1.8%)
Total	29 (100.0%)	29 (100.0%)	663 (100.0%)	662 (100.0%)

Note. Counts are presented as raw numbers with respective valid proportions of the total participants in parentheses.

Teachers who selected “Other” to indicate that teachers are (or should be) evaluated primarily for a reason not provided as a response option in the survey had an opportunity to explain further. Some noted that the evaluation system is multi-purpose (i.e., some combination of professional growth, accountability, state policy compliance, etc.). Other teachers argued that experienced practitioners should not be evaluated annually. For example, one teacher explained that, “I do not think teachers should be formally evaluated every school year. A great teacher is always learning and improving. Administrators soon know those teachers.” Another teacher agreed: “After a number of

good evaluations, teachers should be treated as professionals and not evaluated every year if there are no concerns about them.”

Although certainly in the minority, a few teachers specifically cited other concerns. For example, one teacher wrote that the purpose of the system is “to give administrators power to harass teachers they feel they don’t want on their campus.” Another cited financial motives, explaining that the purpose is “to weed out ‘OLDER’ teachers and bring in the young for less money.” Although these concerns were not frequently expressed, however, as stated prior I sought to represent the voices of teachers through their own words without stifling those with divergent views.

Research Question 2: Intended Implementation

Transparency. When conducting interviews with school administrators and teachers to inform survey development, I sought to better understand their perceptions of the teacher evaluation system design and implementation processes. Interview questions were intended to gauge the perceived transparency of these district-level processes from the viewpoints of both administrators and teachers as subjects of a state policy-directed, but locally-developed and implemented evaluation system. In terms of transparency in system development, administrators described their respective experiences. One administrator explained:

For me it’s been really transparent, and I feel like I’ve known the steps all along the way, especially with the rollout [where the district kept asking for] input... for the rollout of the CES system and how we can better that system and what administrators are looking for and what teachers are looking for to improve on that. I think [that] has been very transparent.

Another administrator agreed, noting that, “I think the district has tried to be very, very transparent.” He or she elaborated further, explaining that “I think the district’s done a very good job at rolling out, providing information, having talks, having quorums for people to come meet. It’s just [that] I don’t know how many [teachers] have taken advantage of that.” A third administrator described a similar experience: “I think it’s been pretty transparent because any time that something’s been changed, we always get information regarding that as far as updates and what’s going to happen.” Based on their interview responses, these administrators seem to have received adequate information about the evaluation system.

However, teachers as a group described varying levels of knowledge about, exposure to, and/or involvement in system design and implementation. One teacher described clear communication from district administration:

The process has been pretty transparent...[in that] there were people from [the] district that came out and explained to us...[a few years ago] when it was first discussed of the 67 percent and the 33 percent, and so I feel like I was fully informed of what was going to happen.

Other teachers described different experiences, especially when asked whether they believed teachers had input into the process. One probationary teacher (i.e., with less than four years of experience in the district) specifically addressed his or her lack of awareness regarding opportunities for teacher input:

I wouldn't say [it has been] transparent.... I know that different teachers were taken from our school and other schools to have a say within the process of it. We

weren't aware that teachers were being pulled into that creation until after the fact.

An experienced teacher expressed his or her initial reaction to the new system:

I don't think it is transparent. I have no idea how this system was put together.

The first time I recognized it, I was shocked because for the 25 years I've been teaching there've always been written notes, there's been some kind of personal input into it. Yes, you need statistics and—for evaluation purposes, I understand that. I really think teachers appreciate some kind of personalization that you're a human being and just being a number is, in my opinion, insulting. It's robotic.

As evidenced by interview responses regarding transparency, school administrators and teachers expressed varying levels of understanding regarding the system components and steps of the evaluation process, as well.

Understanding system components. School administrator interviewees frequently cited their extensive training as helpful in gaining an understanding of the evaluation system components (i.e., professional practice and value-added measures). Referencing the Danielson FFT, one administrator noted that, “you have different indicators, different components of every domain. I think that part of it is easily understood.” Another administrator addressed teachers’ understanding of the Danielson FFT:

I don't know how much teachers really understand about the evidence piece of it. If the evidence isn't there, it isn't there; however, it is a learning process as far as once they receive a particular rating, then that's when it becomes important for them to look at the rubric and look at the examples, that type of thing.

However, another administrator disagreed, explaining:

As far as the process, to me, it's pretty clear. I've not had any teachers say, 'I don't understand. What am I supposed to be doing?' They understand how certain things belong in certain domains. That part, I think, is pretty clear to them.

As evidenced by their interview responses, these administrators held somewhat dichotomous beliefs about teachers' understanding of system components.

Perhaps unsurprisingly, teacher interviewees also reported varying levels of understanding. Some teachers credited their administrator(s) for clarifying the system components. One teacher illustrated this point: "I think they're very clearly defined. My principal's gone over exactly what falls within Danielson's Framework and each of the domains, and we've extensively gone over them last year and this year." He or she continued, explaining that "everything is set out [and] very clear-cut for what's expected of us." Another more experienced teacher cited his or her artifact binder as an organizational tool (i.e., required by some principals in the district but not all), adding that, "mine is pretty extensive. See, I have each domain defined, so it's easy for me to look at it." Another probationary teacher described his or her confusion in classifying evidence in each domain of the Danielson FFT:

If you're new to the system, [or if] you're an administrator that's new, like ours is, you have to learn this. That's hard. This is very difficult to do.... What is it missing? It's confusing, the four domains, [and] where things belong.

Based on their responses, administrator interviewees reported a greater level of understanding with regards to the Danielson FFT than some teachers.

Despite variation in their own understanding of the professional practice component, administrator and teacher interviewees expressed similar confusion about and/or distrust of the value-added model and resulting overall effectiveness classification. When asked about the inclusion of student growth as a system component, one administrator admitted:

I don't know the math on that. I think everyone is—I think there may not be a clear understanding of what that is. I think people understand there's a formula....

I think maybe a clear explanation of how that formula works might help.

Another administrator expressed similar discomfort when explaining the value-added model to teachers, stating that, "I really can't explain it to them other than there's a formula, it's a magic formula that is put into place." He or she also noted teachers' confusion and, in some cases, resulting distrust, writing:

I don't think they [teachers] understand how it works. I think they just know that there's a formula that takes all of that into account and supposedly equalizes the playing field.... Most people are going to trust, and that's good. [But for others, there is] confusion [and] misunderstanding. Then that, of course, leads to a little bit of mistrust. What are they really doing? What is this formula?

A third administrator expressed his or her concern as well, noting that, "as far as the value added piece of it, I know I don't have as good of understanding of it as some people who understand the statistics" but then added that "I do have a pretty good idea of how it comes about and what they use for that." With regards to the resulting overall effectiveness classifications (i.e., Highly Effective, Effective, Developing, and

Ineffective), the same administrator directly addressed the impact of labeling individual teachers:

I know there have been some questions, and it kind of makes me wonder too how they're setting the guidelines of highly effective, effective and so on.

Mathematically it makes sense. You just kind of wonder what that means to the actual person...you almost want to put yourself in that person's position and how they're going to perceive that information.

Teachers, especially those who were relatively new to the district, described feeling confused and overwhelmed. One first-year teacher explained his or her orientation to the evaluation system components, specifically the value-added measure, writing that, "I just had no idea what was going on.... It was very fast and thrown at you, and I didn't really understand it, but I didn't want to feel stupid, you know?" He or she added:

It was very confusing to me. I've been teaching a long time, like I said, but it was so different out here. I was like, 'What? Don't they just come watch me?' I felt kind of silly. Nobody ever told me.

To better contextualize school administrator and teacher understandings, the survey included questions relating to these systems components (i.e., the Danielson FFT and value-added model) as well as the performance group assignment and overall effectiveness classification.

Given a set of Likert-type items used to measure degree of system component understanding, survey participants validated open-ended interview responses. Survey participants indicated their level of agreement when asked if they would be very comfortable explaining each component to a non-educator. Responses suggested that

school administrators and teachers were most comfortable explaining the professional practice score calculation ($M = 3.33$, $SD = 0.78$ and $M = 2.58$, $SD = 0.87$, respectively) and least comfortable explaining the value added score calculation ($M = 2.74$, $SD = 0.59$ and $M = 2.25$, $SD = 0.84$, respectively). See Table 6 for all other system components that school administrators and teachers reported, in order of greatest to least in terms of comfort with corresponding means and standards deviations.

Table 6

Understanding of Teacher Evaluation System Component Calculations

I would be very comfortable explaining to a non-educator how this component is calculated:

Component	<i>n</i>	<i>M</i>	<i>SD</i>
Professional Practice Score			
Administrators	27	3.33	0.78
Teachers	524	2.58	0.87
Overall Effectiveness Classification			
Administrators	27	3.22	0.58
Teachers	521	2.47	0.86
Performance Group Assignment			
Administrators	26	3.00	0.69
Teachers	519	2.36	0.88
Value-added Score			
Administrators	27	2.74	0.59
Teachers	521	2.25	0.84

Note. Likert items were scaled as follows: strongly agree = 4, agree = 3, disagree = 2, and strongly disagree = 1.

Understanding evaluation processes. School administrators and teachers interviewees were reportedly more comfortable with the steps of the evaluation process.

One administrator explained that:

I feel like we've had a lot of training in it [the evaluation process]. You know going through the beginning steps of the trainings, I think there's a lot and so that's really helpful to be prepared for it. It's a tough process...you see there's so much and being able to get all of the information for a certain domain, I think, is kind of tricky, to be honest, in some points. Like I said, just sitting down and talking to the teachers and seeing what they do for those things has been valuable.

Another teacher expressed a similar sentiment, noting that, "as far as the steps go, I think they're very clear-cut, and they're very [well] understood within our staff." Another teacher agreed, explaining:

I think that within our school, our principal went through [the evaluation process in] great detail with us in small groups. We were able to ask questions, and we had a complete understanding of where they take the different [pieces of] information from and how it was created. I think [we] are definitely well-educated. I can't speak for other schools.

Acknowledgement by these interviewees that there may have been variation across schools in terms of communication and utilization of district-provided resources to increase teacher understanding prompted the inclusion of specific, related survey questions.

School administrators and teachers were asked to evaluate the adequacy of district communication efforts as well as the utility of teacher evaluation system resources in helping to increase their understanding of the evaluation purpose, design, and processes. Nearly three-fourths of administrators ($n = 20/27$, 74.1%) reported in the survey that the district communicated *very well* with them regarding system development and

implementation (see Table 7). Although only 22.8% of teachers ($n = 129/565$) selected the same response option, an additional 61.9% ($n = 350/565$) described district communication efforts as *adequate*. Of concern, 15.2% of teachers ($n = 86/565$) reportedly believed that the district did not communicate very well on this topic.

Table 7

Adequacy of District Communication

The District communicated ____ regarding system development and implementation.

	Very Well	Adequately	Not Very Well	<i>n</i>
Administrators	20 (74.1%)	7 (25.9%)		27 (100.0%)
Teachers	129 (22.8%)	350 (61.9%)	86 (15.2%)	565 (100.0%)

Note. Counts are presented as raw numbers with respective valid proportions of the total participants in parentheses.

In response to a series of Likert-type items, school administrators and teachers appraised the utility of various online and professional development resources as well as communication with others both at the district and school levels. School administrators cited communication with their school-level peers as the most helpful resource ($M = 2.69$, $SD = 0.47$) followed by both professional development training led by district administrators ($M = 2.63$, $SD = 0.49$) and Comprehensive Evaluation System (CES) resources ($M = 2.63$, $SD = 0.49$). Teachers reported that formal or informal communication with and professional development led by their school administrators were most helpful ($M = 2.12$, $SD = 0.62$ and $M = 2.10$, $SD = 0.65$, respectively). Although teacher survey respondents reportedly found professional development led by and/or discussions with their school administrators helpful, teacher interviewees

suggested that not all teachers across the district may have had the same opportunities or positive experiences.

Teachers were also asked to appraise the helpfulness of their peers. Interestingly, they cited other teachers at their school ($M = 2.08$, $SD = 0.63$) as more helpful than those who would have had additional training on the teacher evaluation system (e.g., members of the Teacher Evaluation Committee or teachers union representatives). Although most schools have an onsite representative from the committee, teachers union, or both, teachers may have utilized their other peers for any number of reasons (e.g., accessibility, familiarity, level of trust, etc.). See school administrator and teacher appraisals of resource utility organized by type (i.e., online, professional development, and communication) with corresponding means and standard deviations in Table 8.

Table 8

Helpfulness of Online Resources, Professional Development, and Communication with Others

Resources	<i>n</i>	<i>M</i>	<i>SD</i>
Resource Links on District Website			
Administrators	25	2.28	0.61
Teachers	433	1.93	0.62
Videos on District Website			
Administrators	25	2.44	0.58
Teachers	455	1.87	0.63
Comprehensive Evaluation System (CES) Resources			
Administrators	27	2.63	0.49
Teachers	513	1.99	0.61
District Teacher Evaluation Handbook			
Administrators	26	2.35	0.69
Teachers	456	1.91	0.61
Professional Development Led by District Administrators			
Administrators	27	2.63	0.49
Teachers	523	1.95	0.63
Professional Development Led by School Site Leadership			
Administrators	24	2.50	0.51
Teachers	557	2.10	0.65
District Administrators			
Administrators	27	2.56	0.51
Teachers	443	1.94	0.62
School Administrators			
Administrators	26	2.69	0.47
Teachers	536	2.12	0.62
Member of the Teacher Evaluation Committee			
Administrators	20	2.15	0.75
Teachers	312	1.85	0.64
Teachers			
Administrators	25	2.24	0.66
Teachers	496	2.08	0.63
District Education Association (Teachers Union)			
Administrators	-	-	-
Teachers	341	1.87	0.65

Note. Likert items were scaled as follows: very helpful = 3, somewhat helpful = 2, and not very helpful = 1. Responses that the resource was “never accessed” were removed for analyses.

Confusion among teachers regarding evaluation system components and processes may have been unknowingly and unintentionally perpetuated by peers who did not have a clear understanding. For example, regarding the 2012-2013 Teacher Effectiveness Classification Report (a pilot year report received by teachers and viewed by administrators one week before the survey), administrators expressed stronger levels of agreement when asked whether the report was comprehensive and easy to understand and whether the additional resource links were helpful. Teachers expressed greater concern with lingering questions even after reading the report (see Table 9). Responses suggested that perhaps administrators had a better understanding of the components or processes than teachers before reading the report or that the report better met their needs.

Table 9

2012-2013 District Teacher Effectiveness Classification Report

Statement	<i>n</i>	<i>M</i>	<i>SD</i>
Report included all the important information about the teacher evaluation.			
Administrators	27	3.22	0.58
Teachers	489	2.79	0.72
Descriptions for each section of the report were easy to understand.			
Administrators	27	3.07	0.68
Teachers	491	2.58	0.81
Additional resource links helped me better understand the teacher evaluation.			
Administrators	27	3.15	0.46
Teachers	460	2.57	0.78
I still had questions about the evaluation after reading the report.			
Administrators	26	2.23	0.65
Teachers	482	2.82	0.79

Note. Likert items were scaled as follows: strongly agree = 4, agree = 3, disagree = 2, and strongly disagree = 1. Responses that participants had “never seen” the report were removed for analyses.

Research Question 2: Actual Implementation

Fidelity of evaluation processes. School administrators and teachers reported similar rates of completion or participation when provided identical lists of evaluation activities in a closed-ended survey question. For each evaluation activity, at least 90% of administrators ($n = 27$) reported that all/nearly all of the teachers in their school completed/participated in the prescribed evaluation activities during the 2013-2014 school year (see Table 10).

Table 10

Administrators Reported the Proportion of Teachers Who Completed or Participated in Evaluation Activities in the 2013-2014 School Year

Evaluation Activity	All/Nearly All	Some	Few/None	<i>n</i>
Personal Self-Assessment	25 (92.6%)	2 (7.4%)		27 (100.0%)
Individual Prof. Development Plan	27 (100.0%)			27 (100.0%)
Beginning of the Year Conference	27 (100.0%)			27 (100.0%)
Walk-through Observation(s)	26 (96.3%)	1 (3.7%)		27 (100.0%)
Informal Observation(s)	26 (96.3%)	1 (3.7%)		27 (100.0%)
Pre-Conference(s)	27 (100.0%)			27 (100.0%)
Formal Observation(s)	27 (100.0%)			27 (100.0%)
Reflection/Formal Observation(s)	27 (100.0%)			27 (100.0%)
Post Conference(s)	27 (100.0%)			27 (100.0%)
End of Year Conference	24 (92.3%)	2 (7.7)		26 (100.0%)

Note. Counts are presented as raw numbers with respective valid proportions of the total participants in parentheses. End of Year Conferences were in progress at most schools at the time of survey administration.

The vast majority of teachers reported similar rates of participation with the exception of the end of the year conference (some of which took place during or after survey administration) (see Table 11).

Table 11

Teachers Who Reported Completing or Participating in Evaluation Activities in the 2013-2014 School Year

Evaluation Activity	<i>n</i> = 591
Personal Self-Assessment	557 (94.2%)
Individual Professional Development Plan	529 (89.5%)
Beginning of the Year Conference	522 (88.3%)
Walk-through Observation(s)	549 (92.9%)
Informal Observation(s)	552 (93.4%)
Pre-Conference(s)	579 (98.0%)
Formal Observation(s)	587 (99.3%)
Reflection on Formal Observation(s)	563 (95.3%)
Post Conference(s)	574 (97.1%)
End of Year Conference	502 (84.9%)

Note. Counts are presented as raw numbers with respective valid proportions of the total participants in parentheses.

Although fidelity of evaluation processes in terms of steps or activities did not seem problematic across the district, the utility of each activity for teachers was also of interest. School administrators and teachers both reported that the formal classroom observation(s) were the most useful for improving teachers' professional practice ($M = 2.96$, $SD = 0.20$ and $M = 2.57$, $SD = 0.61$, respectively). Despite this initial agreement, teachers reported lower levels of utility than administrators for all activities (with the exception of the personal self-assessment completed at the beginning of the year and the end of year conference). Interestingly, teachers also highly rated the preconference(s) and reflection(s) during the post-conference(s) in terms of utility while administrators cited the walkthroughs (five to fifteen minutes) and informal observations (more than 15 minutes) as the next most useful. In general, teachers reportedly valued opportunities to receive formal feedback from their administrator as opposed to additional and relatively

short observations. See the utility of evaluation activities in order of completion with corresponding means and standard deviations in Table 12.

Table 12

Utility of Evaluation Activities for Improving Teacher Professional Practice

Statement	<i>n</i>	<i>M</i>	<i>SD</i>
Personal Self-Assessment			
Administrators	27	2.19	0.56
Teachers	563	2.30	0.70
Individual Professional Development Plan			
Administrators	27	2.37	0.63
Teachers	550	2.15	0.74
Beginning of the Year Conference			
Administrators	27	2.59	0.57
Teachers	540	2.30	0.71
Walk-through Observation(s)			
Administrators	26	2.88	0.33
Teachers	559	2.35	0.73
Informal Observation(s)			
Administrators	26	2.92	0.27
Teachers	562	2.38	0.72
Pre-Conference(s)			
Administrators	26	2.65	0.49
Teachers	569	2.39	0.68
Formal Observation(s)			
Administrators	26	2.96	0.20
Teachers	571	2.57	0.61
Reflection on Formal Observation(s) during Post Conference(s)			
Administrators	27	2.81	0.40
Teachers	562	2.52	0.65
End of Year Conference			
Administrators	25	2.36	0.70
Teachers	395	2.38	0.70

Note. Likert items were scaled as follows: very useful = 3, somewhat useful = 2, and not very useful = 1. Responses that the activity was “not conducted” were removed for analyses. End of Year Conferences were in progress at most school at the time of survey administration.

Evaluator training and objectivity. Based on completion of/participation in prescribed evaluation activities, fidelity of evaluation processes across the district did not seem problematic. However, when asked to evaluate other specific aspects of the evaluation process on a series of Likert-type items (e.g., evaluator training, objectivity/bias, time spent in the classroom), school administrators and teachers often expressed different views (see Table 13). All administrators ($n = 28/28$, 100.0%, $M = 3.50$, $SD = 0.51$) and the vast majority of teachers ($n = 517/578$, 89.5%, $M = 3.22$, $SD = 0.70$) agreed or strongly agreed that they/their administrators were well trained in the Danielson FFT; however, perhaps unsurprisingly, administrators were more confident than teachers were that they could evaluate objectively and without bias ($M = 3.64$, $SD = 0.49$ and $M = 3.04$, $SD = 0.82$, respectively). (Concerns regarding the adequacy of time spent in the classroom will be discussed in the next section.)

Table 13

Evaluator Training, Objectivity, and Time Spent in the Classroom

I/Administrators am/are able to:

Statement	<i>n</i>	<i>M</i>	<i>SD</i>
Well trained in the Danielson Rubrics to evaluate teachers			
Administrators	28	3.50	0.51
Teachers	578	3.22	0.70
Evaluate teachers objectively and without bias			
Administrators	28	3.64	0.49
Teachers	578	3.04	0.82
Spend enough time in teachers' classrooms to adequately evaluate them			
Administrators	28	3.18	0.55
Teachers	576	2.91	0.92

Note. Likert items were scaled as follows: strongly agree = 4, agree = 3, disagree = 2, and strongly disagree = 1.

Teachers elaborated on their concerns about evaluator subjectivity, describing personal experiences during the evaluation process. One teacher wrote, “The administrator who conducted my evaluation was very subjective and based [my] evaluation on bias towards [me as] a person and not objective data.” Another teacher expressed similar frustration: “Unfortunately, due to having an administrator that is biased, I had a difficult time this year. I feel that I had to take [matters] into my own hands and deal with the inequities of my evaluation.” Another teacher questioned his or her evaluation on these grounds, writing that:

I feel that the final outcome of my teacher score was influenced by my administrator and the negative feelings they [sic] had towards me. That being said, I feel it is important if a teacher is going to be labeled there needs to be some unbiased evaluations involved.

Some teachers described the impact of perceived subjectivity/bias on the part of only one evaluator at their school. For example, one teacher explained:

The evaluation needs to be done by BOTH administrators if there are two at a building. I felt that there was [a] VERY BIG discrepancy in evaluations due to personal bias from different observers. What one person believes is basic may be what another person feels is proficient.

Citing the need for his or her evaluator to spend more time in the classroom to mitigate personal bias, another teacher explained that:

No true amount of training will fix the errors conducted by humans when evaluating others. As humans, we are flawed and [have] personal bias and

opinions will always stir issues due to little to no actual 'time' spent in the classroom.

This teacher's concern about his or her administrator's lack of time spent in the classroom did not appear to be an isolated problem.

Time spent in the classroom. When asked whether teachers' evaluators had spent enough time in their classroom during the 2013-2014 school year, 28.3% of teachers ($n = 163/576$) disagreed or strongly disagreed. Only 7.1% of administrators ($n = 2/28$) reportedly had the same concern. One teacher summarized his or her concern, noting that "administrators do not spend enough time in the classrooms of their teachers and requirements are not unbiased." Some administrators also acknowledged that a significant amount of time is required to properly evaluate teachers although none indicated that the time currently spent was inadequate. For example, one administrator wrote:

I only say that my stress/apprehension has been raised because these scores are such a heavy label for the teachers. I feel a tremendous responsibility to give enough time and opportunity to the teachers in order to truly use the system for their benefit rather than a negative tool.

Another administrator described the evaluation process as limiting time for other administrative duties, noting:

This is a very lengthy process that takes up the majority of the instructional days. The weekends are spent on completing the documents. Even though this benefits the teachers, it is detrimental with other aspects of an administrator's job responsibilities.

Although administrators did not identify a need to spend more time conducting evaluation activities, they certainly acknowledged that a significant amount of time is required to adequately evaluate teachers.

These differing perspectives on the issue of time adequacy in the evaluation process exemplified an emerging trend in the data that will be discussed in detail in Chapter 6. Specifically, administrators and teachers often recognized the importance of the same aspect(s) of the evaluation system or processes (in this case, time spent in classrooms); however, often only one group characterized this aspect as problematic or an area in need of improvement. Even if a common problem were identified, administrators and teachers did not necessarily propose comparable solutions.

Research Question 3: Measuring System Effectiveness

Validity. When measuring the effectiveness of school reform, in this case a teacher evaluation system, Cuban (1998) suggested that the standards used would depend upon who was taking the measurements. In an effort to better understand the perceptions of school administrators and teachers regarding the standards they use for this purpose, I included in the interview and survey instruments series of questions related to the validity, reliability, and fairness of the system (i.e., design, components, implementation, etc.). Perhaps unsurprisingly, perceptions within and across groups varied, although not in the ways that I would have expected.

Content-related validity. To examine evidence of various types of validity (i.e., content, criterion, construct, and consequential), I included series of questions in the interview and survey instruments about the Danielson FFT and value-added model. Interviewees were first asked to assess the content-related validity of the Danielson FFT,

specifically whether the Framework measures the most important aspects of teacher quality, and if not, what domains and/or components were missing. In general, school administrators indicated that the Danielson FFT included the components necessary to adequately measure teacher quality. One administrator explained that, “I think it covers every domain...you have your planning domain, and you have your instruction domain, and really two of those. I think everything’s just as important to the whole teacher package.” Another administrator dismissed the notion that an additional domain should be added, explaining that, “I can’t imagine another domain; that would be horrendous. I mean it’s about as big as it can get and be effective unless you’re going to give everybody another administrator just to run the evaluations.” A few administrators cited redundancy or overlap in the domains although this was not necessarily seen as problematic. One administrator noted that, “I think there’s a lot of overlap, but I think you have to have that because we don’t do things in isolation.” Based on their responses, these interviewees felt confident that the Danielson FFT domains and components adequately measure teacher quality.

The teacher interviewees generally agreed with the administrators in terms of the comprehensiveness of the Danielson FFT. One teacher expressed this sentiment:

I think within all of the domains—it covers the planning and the preparation. It covers classroom organization, classroom management, the knowledge of the students, the contributions you make as a learner—as an educator—to the school, to the community, and things that you do to better yourself. I think those are all components that you need to look at the quality of a teacher.

Another teacher agreed noting that, “I do think that it [the Danielson FFT] measures what is important to be a good teacher.” A third teacher explained that, “I feel like it fully encompasses everything that an effective teacher should be and have.” These responses prompted the inclusion of similar questions on the survey instruments.

School administrator and teacher survey respondents also reportedly believed that the Danielson FFT includes the most important components. More than 90% of administrators ($n = 27/29$, 93.1%) and 73.1% of teachers ($n = 465/636$) indicated that the Framework included all or most of the important characteristics of an effective teacher (see Table 14). When asked specifically what, if any, attributes or characteristics of effective teaching were missing, teacher survey respondents most frequently cited affective attributes such as collaboration with peers/colleagues, rapport with students, teachers’ willingness to accept additional responsibilities, and contributions to the school community. In addition, teachers suggested that examples of what good/effective practices look and sound like at each level of proficiency on the Danielson FFT (i.e., Unsatisfactory, Basic, Proficient, and Distinguished) would be helpful. Some teachers also indicated that the Danielson FFT was misaligned with the professional practices of special education and special area teachers.

Table 14

Danielson Framework for Teaching (FFT)

Danielson FFT includes the important characteristics of an effective teacher.

	All/Most	Some	Only a Few	<i>n</i>
Administrators	27 (93.1%)	2 (6.9%)		29 (100.0%)
Teachers	465 (73.1%)	154 (24.2%)	17 (2.7%)	636 (100.0%)

Note. Counts are presented as raw numbers with respective valid proportions of the total participants in parentheses.

Teacher survey respondents strongly expressed their disagreement, however, with the use of students' scores on AIMS reading, mathematics, and/or science tests in the value-added model to estimate teacher effects, especially for those who teach non-tested grade levels or content areas. One teacher emphatically expressed this concern arguing that, "student achievement is not a valid measure of a teacher's effectiveness." Another cautioned that "student achievement is a snapshot [and reiterated] that one test shouldn't be used to evaluate anyone." Although the district utilizes up to three years of student achievement data, another teacher expressed his or her frustration with the perceived use of test scores as a snapshot measure of student content mastery as well as the lack of student accountability for achievement:

Basing a student's achievement off of one test is not a good indicator of that student's achievement. Also a teacher is penalized for having a class with either 'high' or 'low' students. Many students do nothing but fill in bubbles, yet teachers are held accountable.

Teachers also frequently cited the use of school-level value-added scores for Group B teachers as problematic. One teacher wrote that, "this is hard because the data being used is not ours." Another teacher agreed, explaining his or her frustration as a Group B teacher: "It is discouraging to Group B teachers who are evaluated based upon AIMS scores for students when they don't administer or have students who take the AIMS test." Primary grade teachers (PreK-2) frequently cited the use of grade 3 AIMS scores in their evaluation as troubling. A teacher explained that, "using only student AIMS data instead of authentic assessment isn't necessarily reflective of student achievement. For primary grades not taking AIMS, using those results seems very

disconnected to teacher effectiveness.” Although they were in the minority, it is important to note that a few teachers specifically cited their value-added score as the only valid, reliable measure of their professional performance. One teacher strongly expressed this belief: “The Danielson rating is extremely subjective to the evaluator, whereas student achievement/growth is undeniable.”

When asked whether other important attributes/characteristics of good/effective teaching should be added to the evaluation system, relatively low proportions of administrators and teachers supported adding specific non-test measures (see Table 15). Only 14.3% of administrators ($n = 4/28$) and 10.6% of teachers ($n = 63/595$) agreed that parent satisfaction indicators should be added to the system. Although 17.8% of teachers ($n = 106/597$) suggested that indicators of student attitude, satisfaction, and connection with the teacher and/or school would be appropriate to add, very few administrators agreed ($n = 2/28$, 7.1%). Fewer administrators than teachers also supported the addition of peer- or teacher-based feedback ($n = 1/28$, 3.6% and $n = 102/582$, 17.5%, respectively).

Table 15

Considering Whether the District Should Add Non-test Information to the Teacher Evaluation System

Statement	Yes	Possibly	No	<i>n</i>
Parent satisfaction with teacher/school				
Administrators	4 (14.3%)	8 (28.6%)	16 (57.1%)	28 (100.0%)
Teachers	63 (10.6%)	206 (34.6%)	326 (54.8%)	595 (100.0%)
Student attitude, satisfaction, and connection with teacher/school				
Administrators	2 (7.1%)	10 (35.7%)	16 (57.1%)	28 (100.0%)
Teachers	106 (17.8%)	228 (38.2%)	263 (44.1%)	597 (100.0%)
Peer-based feedback on teacher/school quality				
Administrators	1 (3.6%)	11 (39.3%)	16 (57.1%)	28 (100.0%)
Teachers	102 (17.5%)	263 (45.2%)	217 (37.3%)	582 (100.0%)
Other				
Administrators		1 (20.0%)	4 (80.0%)	5 (100.0%)
Teachers	25 (25.0%)	11 (11.0%)	64 (64.0%)	100 (100.0%)

Note. Counts are presented as raw numbers with respective valid proportions of the total participants in parentheses. Respondents who selected "Other" non-test information subsequently explained their responses.

In general, support for adding alternative measures of student achievement was surprisingly low (see Table 16). Less than one-fourth of administrators and teachers suggested that other assessments (e.g., district benchmark, formative, end-of-course, or other subject [e.g., science, social studies, fine arts, etc.] assessments) should be added. Although 30.5% of teachers ($n = 181/593$) suggested that performance-based assessments would be appropriate, only 18.5% of administrators ($n = 5/27$) agreed. Frequent criticisms of these alternative student achievement or learning measures by both administrators and teachers included: issues of content-validity, reliability, test security, and for some teachers, resistance to any additional testing that could reduce instructional time.

Table 16

Considering Whether the District Should Add Alternative Student Achievement or Learning Measures to the Teacher Evaluation System

Statement	Yes	Possibly	No	<i>n</i>
District Benchmark Assessments				
Administrators	5 (17.9%)	18 (64.3%)	5 (17.9%)	28 (100.0%)
Teachers	110 (18.6%)	258 (43.7%)	222 (37.6%)	590 (100.0%)
Formative Measures				
Administrators	2 (7.4%)	13 (48.1%)	12 (44.4%)	27 (100.0%)
Teachers	121 (20.6%)	294 (50.1%)	172 (29.3%)	587 (100.0%)
Performance-Based Assessments				
Administrators	5 (18.5%)	8 (29.6%)	14 (51.9%)	27 (100.0%)
Teachers	181 (30.5%)	270 (45.5%)	142 (23.9%)	593 (100.0%)
End-of-Course Assessments				
Administrators	5 (18.5%)	14 (51.9%)	8 (29.6%)	27 (100.0%)
Teachers	101 (17.3%)	280 (47.9%)	203 (34.8%)	584 (100.0%)
Course Grades or Grade Point Average (GPA)				
Administrators	1 (3.8%)	9 (34.6%)	16 (61.5%)	26 (100.0%)
Teachers	73 (12.5%)	213 (36.5%)	298 (51.0%)	584 (100.0%)
Other Subject Area Assessments				
Administrators	4 (16.0%)	11 (44.0%)	10 (40.0%)	25 (100.0%)
Teachers	131 (22.7%)	230 (39.8%)	217 (37.5%)	578 (100.0%)
School College-Ready Indicators				
Administrators	2 (7.4%)	12 (44.4%)	13 (48.1%)	27 (100.0%)
Teachers	60 (10.4%)	198 (34.3%)	319 (55.3%)	577 (100.0%)
School Graduation or Dropout Rates				
Administrators	2 (8.7%)	8 (34.8%)	13 (56.5%)	23 (100.0%)
Teachers	47 (9.0%)	149 (28.4%)	328 (62.6%)	524 (100.0%)
Other				
Administrators			5 (100.0%)	5 (100.0%)
Teachers	17 (13.4%)	18 (14.2%)	92 (72.4%)	127 (100.0%)

Note. Counts are presented as raw numbers with respective valid proportions of the total participants in parentheses. “Other Subject Areas” could include science, social studies, etc. “School College-Ready Indicators” could include college-readiness assessments administered to students in grades 7-8. Respondents who selected “Other” alternative student achievement/learning measures subsequently explained their responses.

Criterion-related validity. In terms of criterion-related validity, interviewees were also asked whether they would place their confidence in a teacher’s professional practice

score (on the Danielson FFT) or value-added score if the two were not aligned. School administrators generally expressed confidence in teachers' value-added scores. One administrator questioned whether misalignment of the professional practice and value-added scores was likely to occur but suggested that confidence should be placed in the value-added score: "They should be aligned in my opinion. I think they would be aligned. That is just knowing the data. Let's say the value-added score is higher then we're missing something, obviously, in the professional practice rating." Another administrator agreed, responding in the context of his or her experience as a parent:

I believe that when I have teachers on my campus who have high and low [value-added] scores that the majority of teachers that I would place my kid with bring in AIMS scores that are higher. Those are teachers when I go in [to observe], you feel it in their classroom climate.

Another administrator also suggested that valid, reliable tests should yield value-added scores that are aligned to teachers' professional practice scores. He or she specifically placed confidence in the value-added scores of Group A teachers but acknowledged that other factors may affect that interpretation, explaining: "If it's a Group A teacher...and if you have a good test that's valid and reliable and it's aligned to the curriculum...it can be the most powerful thing. If it's not, there are so many factors there, too." Although these administrators considered the value-added score to be a more valid measure, teacher interviewees expressed dichotomous views on this.

Teachers who favored the value-added score as a better measure of teacher quality also acknowledged that this may not be true for all teachers. For example, one teacher explained that, "I think in general, it'll [the value-added score] will give you a good idea.

I think there's always special circumstances...[but] I feel like if their Danielson score is very, very low, then that's a very big problem." Another teacher agreed, describing the classroom observation as a "show on that one time" and adding that "the [value-added] score is the value part, the best reflection of what you are doing." He or she continued, explaining that, "I'm a math person so I guess, of course, I am going to lean towards statistics."

However, several teachers expressed confidence in the professional practice score. A teacher noted that "I would think that for me, at least, I would go more off my observations in the classroom [rather] than that final [value-added] score." Another teacher explained the importance of an observation to "see" what is really happening in the classroom:

If I can go into that classroom and I can see that students are engaged, and I can see that the teacher is teaching them something that they are supposed to be teaching, and I can tell it's a healthy classroom environment, I'm going to go with the professional score.

The divergent views expressed by teacher interviewees prompted the inclusion of survey questions to determine which measure is perceived as the best indicator of effective teaching and how weighting should be assigned to the measures.

Construct-related validity. When asked which measure (i.e., professional practice, student achievement, a combination, or neither) is the best indicator of effective teaching, school administrator responses did not necessarily reflect interviewee comments (see Table 17). For example, only 6.9% of administrators ($n = 2/29$) reportedly believed that student achievement should be the sole indicator. Rather 82.8% ($n = 24/29$)

supported a combination of professional practice and student achievement measures. In contrast, teachers' survey responses were more reflective of interviewee sentiments. More than half of teachers ($n = 365/637$, 57.3%) indicated that a combination of measures should be used; however, it is important to note that more than one-fourth ($n = 181/637$, 28.4%) believed only professional practice scores should be considered.

Table 17

Best Indicator of Effective Teaching

	Professional Practice	Student Achievement	Combination	Neither	<i>n</i>
Administrators	3 (10.3%)	2 (6.9%)	24 (82.8%)		29 (100.0%)
Teachers	181 (28.4%)	26 (4.1%)	365 (57.3%)	65 (10.2%)	637 (100.0%)

Note. Counts are presented as raw numbers with respective valid proportions of the total participants in parentheses.

The weighting of teachers' value-added and professional practice scores (33.0% and 67.0%, respectively) in the teacher evaluation system was determined by the district in compliance with the Arizona State Board of Education and Senate Bill 1040 (Arizona Revised Statutes §15-203 (A) (38)). However, I sought to better understand the value placed on each measure by administrators and teachers irrespective of policy mandates. As part of the survey, respondents were asked to assign weights to the respective measures (on a scale of 0 to 100% such that weights totaled to 100%). Respondents also had the option to assign weights for up to two "Other" measures and define/describe the measure(s) as an open-ended response. The mean weights (and corresponding standard deviations) assigned to each measure by group are provided in Table 18. Results of t-tests of independence indicate that there was no statistically significant difference between the mean weights assigned by each group to the Danielson FFT or the "Other" measures.

However, there was a statistically significant difference in the mean weight assigned to the student achievement measure (VAM score) by administrators ($M = 34.56$, $SD = 11.30$) and teachers ($M = 27.01$, $SD = 13.95$); $t(547) = 2.76$, $p = 0.006$. In short, school administrators assigned a mean weight of 34.56% to teachers' value-added scores (slightly higher than the current weighting of 33.0%). However, teachers assigned a mean weight of 27.01% to their value-added scores. This mean difference of 7.55% suggests that school administrators on average place greater value on teachers' value-added scores as a measure of effectiveness than teachers do. This finding validates previously discussed interview data.

Table 18

Weighting Assigned by School Administrators and Teachers to Measures of Teacher Effectiveness

	<i>n</i>	<i>M</i>	<i>SD</i>
Danielson FFT			
Administrators	27	62.30	13.38
Teachers	536	67.95	18.97
Student Achievement			
Administrators	27	34.56	11.30
Teachers	522	27.01	13.95
Other (1)			
Administrators	4	16.25	11.09
Teachers	140	23.47	20.80
Other (2)			
Administrators	2	10.00	14.14
Teachers	58	11.93	14.93

Note: Respondents assigned weights to each measure on a scale of 0 to 100. Those who assigned weights to “Other” measures of teacher effectiveness subsequently explained their responses.

Consequential validity. Widespread concerns expressed by interviewees and survey respondents about the consequential validity of the evaluation system necessitated

a survey question related to the perceived representativeness (construct-related validity) of teachers' overall effectiveness classifications (i.e., Highly Effective, Effective, Developing, and Ineffective) as these labels are interpreted and used to make high-stakes decisions. I argue that school administrator (to some extent) and teacher perceptions of the effectiveness classification as (un)representative of professional performance are among the most important study findings. Interestingly, 28.6% of administrators ($n = 8/28$) and 42.2% of teachers ($n = 216/512$) disagreed or strongly disagreed that teachers'/their overall effectiveness classification(s) were/was representative. Again, teachers reported lower confidence in the representativeness of their own high-stakes label than administrators. This is certainly troubling given the accountability policy supposition that valid inferences about professional performance can and should be made from teachers' effectiveness classification labels (see Table 19).

Table 19

Teacher Overall Effectiveness Classification Labels

My/Teacher's Overall Effectiveness Classification Label(s) was/were an accurate representation of my/their professional performance.

	<i>n</i>	<i>M</i>	<i>SD</i>
Administrators	28	2.79	0.69
Teachers	512	2.50	0.95

Note. Likert items were scaled as follows: strongly agree = 4, agree = 3, disagree = 2, and strongly disagree = 1.

Reliability. Administrator and teacher perceptions regarding the reliability of evaluator ratings on the Danielson FFT rubric as well as value-added scores over time merit closer examination, especially given the high-stakes consequences associated with teachers' overall effectiveness classifications. Although teachers, and in some cases

administrators, challenged whether evidence of validity for the Danielson FFT and value-added model are in fact adequate, I discuss reliability in this section under the questionable assumption that adequate evidence exists.

Although I expected that teacher interviewees might question the inter-rater reliability of evaluators in their assignment of ratings, administrators also raised concerns about reliability both within and across schools. An administrator acknowledged this as a potential problem, explaining:

I don't know that there's consistency. I think that would be hard to say. I personally have had conversations with a few people. I feel like maybe some people can be a little more lenient with some [teachers] than others.

Another administrator acknowledged that variation in evaluator experience levels likely contributes to the problem, asking:

How consistent would they be? Well, it depends on defining the term 'consistent.' I think they're inconsistent because consistency means you're doing it all the time at high levels. That's pretty hard to do. We have some very young people out there who are just learning, and we have some assistant principals... [in their] first year.

Administrators also frequently described how they work with the other evaluator(s) at their school to increase inter-rater reliability. One noted that "I think we're still people. There are always going to be some variables, but we have had the conversations, so we know what to look for. I think that is more consistent than it has been." Although this administrator expressed some measure of confidence, another suggested that consistency across schools would be difficult, if not impossible, to achieve:

No, I don't think that's possible. I think there's going to be some consistency, but I think that you'll also need to look at the volume of the size of the district, the number of administrators doing the evaluating.

Although some administrators were skeptical that efforts to increase inter-rater reliability would be successful, most acknowledged that ratings both within their own school and across the district may be unreliable and expressed a need for additional training.

Teacher interviewees frequently implied or, in some cases, directly asserted that the inter-rater reliability of evaluators at their school was unacceptably low. One teacher expressed his or her skepticism, explaining that "I would like to think that everyone is the same or equally evaluated, but I don't think that it's 100 percent true." Another teacher described peers' varying levels of comfort with their respective evaluator as evidence of inconsistency: "I feel like the people who are evaluated by the principal get a little more nervous. The people who get evaluated by the assistant principal seem to be fine with it." When asked whether he or she would expect to receive the same Danielson FFT ratings at another school, a third teacher explained:

I don't know if I would have the same exact rating. I feel that I'm effective, and it has always come out well, and with different administrators. I would be curious [though]...to have a different administrator from another school come in who is completely unbiased and doesn't know [me], and see if they view [me] as effective.

Teacher interviewees questioned evaluator consistency, prompting the inclusion of a series of Likert-type items on the survey gauging their support for implementation measures to increase confidence in the reliability of Danielson ratings.

When asked whether the teacher evaluation system would be significantly improved if teachers were evaluated by more than one observer, an expert in their instructional content area, an external evaluator (outside their school), and/or a peer-evaluator(s) (at least in part), administrators most strongly supported the use of multiple observers ($M = 3.00$, $SD = 0.69$) and content area experts ($M = 2.80$, $SD = 0.58$). Teachers responded similarly, preferring content area experts ($M = 2.91$, $SD = 0.78$) and multiple observers ($M = 2.87$, $SD = 0.79$). See additional proposed implementation measures with corresponding means and standard deviations in Table 20.

Table 20

Improving the Teacher Evaluation System Using Multiple, Content-specific, External, and/or Peer Evaluators

The District Teacher Evaluation System would be significantly improved if teachers were evaluated by...

Statement	<i>n</i>	<i>M</i>	<i>SD</i>
More than one observer			
Administrators	26	3.00	0.69
Teachers	527	2.87	0.79
An expert in their instructional content area			
Administrators	25	2.80	0.58
Teachers	567	2.91	0.78
External evaluators (to their school)			
Administrators	26	2.69	0.88
Teachers	570	2.28	0.86
Peer evaluators (at least in part)			
Administrators	26	2.58	0.86
Teachers	572	2.46	0.83

Note. Likert items were scaled as follows: strongly agree = 4, agree = 3, disagree = 2, and strongly disagree = 1.

In their open-ended survey responses, administrators frequently expressed their support for measures to increase perceivably low inter-rater reliability. An administrator

described the benefits of external evaluators, noting that, “[I] love the idea of someone not on the campus [conducting the evaluation]. It takes the bias (both for and against) out.” Another administrator supported the use of multiple or secondary observers as a professional development activity:

I would like to practice the system in different schools with a partner so that we could discuss the data collected and the classification [rating] we would give to better align my scores to other evaluators. I would like to do this in a variety of settings (Title 1 and non-Title 1 schools).

However, based on their open-ended survey responses, it is important to note that a few administrators did not seem to view inter-rater reliability as a problem. One administrator suggested that teachers’ concerns may be due to a lack of understanding, explaining that, “teachers need to understand more thoroughly how they are evaluated. Some teachers believe their evaluation scores are lower because of administrator differences.” Although this administrator did not necessarily view teachers’ skepticism as indicative of evaluator inconsistency, he or she illustrated another more common theme—the need for additional training.

When asked whether the teacher evaluation system would be significantly improved if additional training on the Danielson FFT rubric were provided to administrators and/or teachers, administrator respondents more strongly supported training for teachers ($M = 3.27$, $SD = 0.72$). Interestingly, teachers agreed ($M = 2.66$, $SD = 0.77$). Administrators and teachers both indicated that additional clarification on Danielson FFT rubric would improve the system as well (see Table 21).

Table 21

Improving the Teacher Evaluation System through Additional Training and/or Danielson Framework for Teaching (FFT) Rubric Clarification

The District Teacher Evaluation System would be significantly improved if:

Statement	<i>n</i>	<i>M</i>	<i>SD</i>
Administrators received more training on the Danielson FFT Rubric			
Administrators	26	2.81	0.69
Teachers	561	2.54	0.73
Teachers received more training on the Danielson FFT Rubric			
Administrators	26	3.27	0.72
Teachers	569	2.66	0.77
The Danielson FFT Rubric was clarified or better defined			
Administrators	25	2.80	0.76
Teachers	568	2.75	0.76

Note. Likert items were scaled as follows: strongly agree = 4, agree = 3, disagree = 2, and strongly disagree = 1.

Although value-added proponents suggest that the use of multiple years of student achievement data increases reliability, teachers frequently criticized this approach for a variety of reasons. For examples, teachers frequently reported that the use of multiple years of AIMS scores is still an inadequate representation of overall student growth as only three content areas (i.e., reading, mathematics, and science) are tested. One teacher noted that, “we need to look at student growth throughout the year in all areas, not just AIMS.” Another teacher described his or her concerns about the instability of and inferences to be made from unreliable value-added scores over multiple years:

Children do not grow at the same rate each year. The student achievement piece implies that all students will grow at the same rate each year and can achieve one year’s growth each year. A child may grow 1/2 a year one at one grade and grow

1 1/2 years the next year. One year the teacher is determined to be bad and the other year the teacher is wonderful, but the results could be developmental only.

A teacher in Group A questioned the calculation of value-added scores over two years for those who change group classifications, noting that he or she will be classified as a Group B teacher next year: “Don't compare apples and oranges. [I am a Group] A teacher becoming a [Group] B the next year and moving from one school to the next. Those two years cannot be compared for growth.” These teachers’ statements reflected the degree of skepticism commonly expressed by their peers regarding the use of students’ AIMS scores in only three content areas to calculate a value-added score each year for every teacher in the district.

Fairness. School administrators and teachers in both interview and survey responses expressed concerns about the fairness of the evaluation system design and/or components (i.e., professional practice and/or value-added measures). In terms of the system’s design, both groups criticized the classification of teachers into either Group A or B. Because Group A teachers (e.g., elementary self-contained classroom teachers in grades 3-6) have achievement data available for their individual students or their content area, they receive a value-added score based on their students’ scores on AIMS reading, mathematics, science, or a combination of these. Teachers for whom this is not the case (e.g., those who teach social studies in grade 7-8, special areas [i.e., art, music, and physical education], etc.) are considered part of Group B and receive a value-added score based on grade- or school-level data.

Administrator and teacher interviewees frequently questioned the validity of this use of student achievement data, especially for teachers in Group B who receive grade- or

school-level value-added scores. One administrator aptly summarized this widely expressed concern:

If we're just trying to manufacture data that would somehow show what this teacher is doing, I think you can do it with a classroom teacher [in Group A]. [However,] it's hard to control all those factors and say this teacher had growth based on this [sic] data. You certainly can't with a [Group] B teacher. I don't think that's really valid and reliable, to say okay, for the whole school, this teacher, this is their growth.

Although administrators cited concerns about group classification in their interview responses, none directly referenced this topic in their survey responses.

In contrast, teacher interviewees and survey respondents in both Groups A and B frequently described the Group B classification as unfair. For example, teachers in grades PreK-2 expressed a common concern, namely the inclusion in their composite teacher evaluation score of grade-level value-added scores for current grade 3 teachers. One teacher's survey statement aptly reflects the sentiment most frequently expressed by his or her peers: "Let's use teacher data from the grade we teach, when we teach it, not the results of AIMS scores 2-3 years after we teach the group." Another teacher explained his or her frustration:

Teachers cannot be judged on how [well] students do for AIMS. I teach Kindergarten so after having different teachers for three years it is not fair for me to get judged on how they do on AIMS. Teachers should be judged solely on their classroom and their end of the year tests [and] not [on] how they [students] do in three years after three different teachers.

When asked whether the evaluation system fairly measures the professional quality of teachers in Group A and in Group B overall, administrators reported higher levels of agreement than teachers (see Table 22). As might be expected based on interview responses, teachers indicated that the system more fairly evaluated their peers in Group A ($M = 2.57, SD = 0.79$) than in Group B ($M = 2.24, SD = 0.88$).

The inability of the evaluation system as a whole to account for student background characteristics was also one of the most frequently expressed concerns among teachers. Demonstrating perceptual differences, teachers ($M = 2.22, SD = 0.85$) reportedly had less confidence than administrators ($M = 2.93, SD = 0.54$) in the system's ability to control for student-level variables.

Table 22

Adequacy and Fairness of the Teacher Evaluation System

The District Teacher Evaluation System...

Statement	<i>n</i>	<i>M</i>	<i>SD</i>
Accurately captures the impact of teachers on student motivation, attitudes, and engagement			
Administrators	28	3.11	0.50
Teachers	580	2.55	0.82
Adequately takes into account the influence of student background characteristics			
Administrators	28	2.93	0.54
Teachers	579	2.22	0.85
Fairly measures the professional quality of teachers in Group A			
Administrators	28	3.07	0.47
Teachers	569	2.57	0.79
Fairly measures the professional quality of teachers in Group B			
Administrators	28	2.54	0.64
Teachers	529	2.24	0.88

Note. Likert items were scaled as follows: strongly agree = 4, agree = 3, disagree = 2, and strongly disagree = 1.

Although this particular survey question did not differentiate between professional practice and value-added measures, respondents directly addressed the likely biasing impact of student background characteristics and other exogenous variables on value-added estimates in their open-ended responses. Teachers specifically cited student background characteristics (e.g., English language proficiency, eligibility for gifted or special education services) as well as classroom dynamics (e.g., student interactions, classroom size) and out-of-school factors (e.g., poverty, level of parental involvement) as impacting student achievement. A teacher survey respondent illustrated this common sentiment, arguing that, “student achievement is not a valid measure of a teacher’s effectiveness. There are too many variables in a student’s life that weigh into what a student retains or can comprehend.” Another teacher summarized his or her concern, writing that “in some respects, teachers should be accountable for student achievement, but there are too many factors that contribute to it that are out of a teacher's control.” Teachers also often identified poverty as impactful, noting that, “teaching in a Title 1 school is very challenging with students coming from low-income families, [and] different cultures. It [the student population] is also very transient.” Another teacher made the following comparison to illustrate the influence of exogenous variables on measures of teacher effectiveness: “Judging teachers on their students’ test scores makes as much sense as judging a farmer on crops without accounting for drought, freezes or diseases.”

Similarly, administrator and teacher perceptions differed in terms of the ability of the evaluation system as a whole to capture the impact of teachers on student motivation, attitudes and engagement. Based on survey results, administrators again expressed greater

confidence in the evaluation system to reflect teachers' impact in these areas (see Table 22). Although some teachers emphasized in their written responses the inadequacy of the Danielson FFT in capturing the positive impact they have on their students, others expressed their frustration, arguing that teachers should not be held accountable when they are unsuccessful in motivating students to put forth their best effort on high-stakes standardized tests. One teacher explained that the system should "help improve the quality of their professional practice to become better teachers/educators and not hold them accountable when scores are not being met due to lack of motivation of their students." Another teacher perceived district culture as partly to blame for a perpetual lack of student accountability:

Unfortunately, I believe that as long as this district allows students to not perform in the classroom without [any] consequence other than their earned grade, it will be difficult for teachers to increase their ratings. The lack of student accountability for learning and growth is unacceptable.

Although this teacher's assertion suggested that the use of student achievement data to evaluate teachers is inherently unfair as a result of low levels of student motivation, it is important to note that teachers challenged system fairness for a myriad of reasons.

Perhaps even more importantly, teachers expressed lower confidence in their ability to control, and improve, their future professional practice score, value-added score, and overall effectiveness classification than administrators (see Table 23).

Teachers were least confident in their ability to impact their future value-added score.

Table 23

Ability to Control and Improve Future Teacher Evaluation System Outcomes

I/Teachers can control, and improve, my/their future:

Statement	<i>n</i>	<i>M</i>	<i>SD</i>
Professional Practice Score			
Administrators	27	3.78	0.42
Teachers	529	3.16	0.78
Overall Effectiveness Classification			
Administrators	27	3.44	0.51
Teachers	527	2.78	0.84
Value-added Score			
Administrators	27	3.22	0.64
Teachers	519	2.50	0.95

Note. Likert items were scaled as follows: strongly agree = 4, agree = 3, disagree = 2, and strongly disagree = 1.

Intended and unintended consequences. These results merit closer examination of the perceived intended and unintended consequences of the teacher evaluation system design and/or implementation. School administrators and teachers largely agreed that the evaluation system should be used to help teachers improve their professional practice. Despite this commonly shared understanding, many teachers were skeptical that the system was actually being used for its stated purpose. Given their divergent views, both groups were asked to discuss the evaluation system in terms of intended and unintended consequences.

Impact on professional practice. Based on interviewee responses, administrators expressed confidence that the system was having a positive impact on their own practice as evaluators as well as that of their teachers. One administrator described his or her experience, explaining:

I think it makes you stronger. I think there's always a learning curve, and every time we go through a process I try and learn something new about it...I think it just makes me a stronger evaluator and a stronger administrator.

When asked about the impact on teachers, the administrator added that, "I think they're more aware. They're more willing to take advice or support. They know what the expectation is, and if they need help, they know we're going to get them whatever they need." Another administrator also cited his or her professional growth, noting that, "it's really deepened my understanding and my knowledge of how we evaluate teachers and also just what makes a good teacher." With regards to the impact on teachers, the administrator continued, "It helps them grow professionally because they're looking at that rubric and taking ownership of [it].... I'm sure a lot of them think it can be pretty intense; however, I do feel like it has helped them be more purposeful in what they're planning...it's not necessarily a game of the dog and pony show in instruction." These responses exemplify the beliefs most frequently expressed by administrator interviewees, namely that they felt more confident and prepared as evaluators, and their teachers seemed to have assumed greater responsibility for their professional practice and evaluation outcomes.

Although teacher interviewees varied to some extent in their responses on this topic, most disagreed that the evaluation process had a generally positive impact or even any impact for that matter on their professional practice. One teacher described his or her reaction to positive feedback from an evaluator but disagreed that the evaluation process had a direct impact on his or her practice, explaining that, "I felt actually kind of happy about it. It really didn't affect me.... It gave me a little boost because I was happy that

they got to come and see, and they could tell I did a good job.” Another teacher replied that, “I’m not sure it’s really affected me in any way at all. I go in, and we talk about it, and I come in here [the classroom], and I do what I’m going to do.... I’ve been very fortunate in that I am apparently good at doing my job.” A third teacher also described the negligible impact of the system on his or her self-confidence as a professional: “It hasn’t impacted me in any way.... I close the door, and I am who I am. I love my kids, and this is who I am. I don’t really care what that [evaluation] says or how many X’s I have in the ‘excellent’ box. I don’t care.” Another teacher described the limited impact on his or her peers, noting that, “I don’t know really that it has impacted teachers that I’ve seen. I know myself—I’m always trying to put my best foot forward and get the A’s.” Another teacher also disregarded whether the evaluation outcomes were impactful for other teachers, noting that, “a lot of people aren’t going to take them seriously. At this school anyway that seems to be the general consensus I’m getting.” These teacher interviewees generally dismissed the evaluation process and outcomes as impacting themselves or their peers.

Survey responses generally validated interview data. When asked how the system had impacted their own practice, 92.3% of administrators ($n = 24/26$) indicated that the impact had been generally positive (see Table 24). Nearly all ($n = 25/26$, 96.2%) believed that the system had the same impact on their teachers. However, only 38.5% of teachers ($n = 218/566$) also described the system impact on their professional practice as generally positive, and, of great concern, more than half of teachers ($n = 348/566$, 61.5%) reported that the system had no real or a generally negative impact in this regard.

Table 24

Impact of the Teacher Evaluation System on Professional Practice

Statement	Generally Positive	No Real Impact	Generally Negative	<i>n</i>
Impact on Administrators				
Administrators	24 (92.3%)	1 (3.8%)	1 (3.8%)	26 (100.0%)
Teachers	-	-	-	-
Impact on Teachers				
Administrators	25 (96.2%)		1 (3.8%)	26 (100.0%)
Teachers	218 (38.5%)	250 (44.2%)	98 (17.3%)	566 (100.0%)

Note. Counts are presented as raw numbers with respective valid proportions of the total participants in parentheses.

In response to a close-ended survey question, administrators and teachers identified which areas of their/teachers' professional practice, if any, had been impacted by the evaluation system. Regarding their own practice, administrators most frequently indicated that the system had created dialogue with their teachers ($n = 23/25$, 92.0%) and provided clarity and focus on good/effective teaching ($n = 22/25$, 88.0%) (see Table 25).

Table 25

Administrators Reported the Impacts of the Teacher Evaluation System on Their Professional Practice

Impact	<i>n</i> = 25
Created dialogue with teachers	23 (92.0%)
Provided clarify and focus on good/effective teaching	22 (88.0%)
Prompted reflection on professional practice	19 (76.0%)
Positive: Narrowed evaluation to the Danielson FFT components	17 (68.0%)
Raised level of stress/apprehension	8 (32.0%)
Increased focus on College and Career Ready Standards	4 (16.0%)
Increased focus on the state standardized assessment	3 (12.0%)
Other	2 (8.0%)
Negative: Narrowed evaluation to the Danielson FFT components	

Note. Counts are presented as raw numbers with respective valid proportions of the total participants in parentheses. Respondents who selected "Other" impacts subsequently explained their responses.

Administrators also cited clarity and focus on good/effective teaching and the creation of dialogue as impactful for teachers. In contrast, teachers most frequently described the system as raising their levels of stress/apprehension ($n = 340/537$, 63.3%) and prompting reflection on their professional practice ($n = 268/537$, 49.9%). Increased clarity and focus on good/effective teaching and creation of dialogue with administrators were the third and fourth most frequently cited by teachers (see Table 26).

Table 26

Impacts of the Teacher Evaluation System on Teacher Professional Practice

Impact	Administrators $n = 26$	Teachers $n = 537$
Provided clarify and focus on good/effective teaching	21 (80.8%)	248 (46.2%)
Created dialogue with school administrators	20 (76.9%)	199 (37.1%)
Prompted reflection on professional practice	20 (76.9%)	268 (49.9%)
Raised level of stress/apprehension	17 (65.4%)	340 (63.3%)
Positive: Narrowed practices to the Danielson FFT components	13 (50.0%)	145 (27.0%)
Enhanced focus on individualized student instruction	8 (30.8%)	146 (27.2%)
Increased use of innovative instructional techniques/activities	6 (23.1%)	173 (32.2%)
Increased focus on the state standardized assessment	3 (11.5%)	114 (21.2%)
Improved communication with parents	2 (7.7%)	76 (14.2%)
Increased focus on College and Career Ready Standards	2 (7.7%)	70 (13.0%)
Other	1 (3.8%)	54 (10.1%)
Reduced use of innovative instructional techniques/activities	1 (3.8%)	52 (9.7%)
Negative: Narrowed practices to the Danielson FFT components		92 (17.1%)

Note. Counts are presented as raw numbers with respective valid proportions of the total participants in parentheses. Respondents who selected "Other" impacts subsequently explained their responses.

Impact on student achievement. Although interviewees were not directly asked to describe the impact of the evaluation system on student academic achievement and learning, I included a survey question on this topic. Based on the sense of indifference expressed by teacher interviewees in terms of the impact of the evaluation system on their own professional practice, I wanted to know whether they believed that the system

impacted their students, especially given current policy debates as to whether stronger teacher accountability systems actually increase student learning. The vast majority of administrators ($n = 21/25$, 84.0%) agreed that the impact on student academic achievement and learning is generally positive (see Table 27). Teachers expressed a markedly different view, however, again. In total, 69.5 % of teachers ($n = 388/558$) believed that the system had no real or a generally negative impact on students in this regard.

Table 27

Impact of the Teacher Evaluation System on Student Academic Achievement and Learning

	Generally Positive	No Real Impact	Generally Negative	<i>n</i>
Administrators	21 (84.0%)	4 (16.0%)		25 (100.0%)
Teachers	170 (30.5%)	314 (56.3%)	74 (13.3%)	558 (100.0%)

Note. Counts are presented as raw numbers with respective valid proportions of the total participants in parentheses.

Impact on teacher hiring and retention. State legislation gives school districts and charter schools in Arizona the right to request an individual teacher’s evaluation report including his or her effectiveness classification from a previous employer (Arizona Revised Statutes §15-537 (I) (3)). Although prospective employers may consider evaluation reports during the hiring process, they are prohibited under state statute from releasing the information to any other person, entity, or school district (Arizona Revised Statutes §15-537 (I) (3)). Although some administrators and, perhaps surprisingly, teachers suggested that the evaluation system should be used to recruit, hire, promote, and retain the most effective teachers, both groups acknowledged the unintended

consequences likely to result from making high-stakes decisions based on evaluation outcomes.

A few administrator interviewees in this study argued that all available information, including prior evaluation results, should be considered when making hiring decisions. It is important to note that AIMS student achievement data are not immediately available after testing, and as a result, teachers do not receive their overall effectiveness classifications/labels until the start of the next school year. Although a teacher's effectiveness classification may not be available for the most recent school year, one administrator explained his or her hesitation in hiring a teacher who had been previously classified as ineffective:

I would think twice before hiring somebody who was ineffective. I think [for a] developing [teacher], if they were in their first three years of teaching, it probably wouldn't bother me. If it was a teacher who had been around for a while but was still developing, that would probably be a deal breaker.

Although this administrator cited the effectiveness classification as a useful indicator to hire the most effective teachers from outside the school and/or district, administrator and teacher interviewees more frequently described increased teacher turnover as problematic.

In fact, several interviewees predicted that retaining effective teachers or even those with strong potential will become increasingly difficult. An administrator predicted that lower than desirable professional practice ratings at the end of the school year may prompt teachers to seek employment in another school or district before overall effectiveness classifications are released: "If it's right before the hiring season starts or

right before contracts come out, it [professional practice ratings] might have an impact.”

The administrator also acknowledged that struggling teachers with strong potential might be even more likely to leave if not given adequate support, adding that “if they don’t have any compensation increase and they’re said to be ineffective when they know they’re busting their tail..., then they might choose to go somewhere else. I wouldn’t blame them.” A third administrator poignantly described the impact of the “Ineffective” label on an individual teacher:

When you label somebody ineffective, that’s very detrimental to [them]. I mean, you might as well slap them in the face and cut their knees off, [and a] whole bunch of [other] stuff that’s horrible. I don’t think labeling someone like that helps them become any better. If it’s anybody at all that you want to keep because you see a lot of potential, you’re not going to be able to keep them.

Administrators specifically anticipated difficulties in retaining younger, less experienced teachers who with adequate support have the potential to be effective or highly effective educators. An administrator described the hypothetical reaction of a first-year teacher, explaining that once “you tell them that they’re inadequate” then they are “going to do something else [as a profession].” He or she cautioned, “That’s why you’ve got to be really careful before you label anything.”

Teacher interviewees reinforced administrators’ concerns about the negative impact of high turnover. Several described strikingly similar scenarios. With regards to the use of effectiveness classifications to make hiring decisions, one teacher admitted that he or she would be more likely to hire an effective teacher, suggesting that, “if a future employer were to pick between two teachers [with] the same credentials, same

experience, but one teacher has an effective [label], and one teacher has a developing [label], I would obviously go with the effective teacher.” However, he or she cautioned against making such an inference based on labels assigned by different districts, arguing that the label “may have had something [more to do with] a scoring difference between the two.” Although this teacher recognized the danger in comparing evaluation classifications across districts, most focused on other unintended consequences associated with labeling teachers in a high-stakes environment.

Teachers frequently cited the risk of increased attrition among both relatively new and more experienced teachers. One teacher explained that, “I understand that the whole purpose about this is to weed out those ineffective teachers, but I still think that people need to take into consideration new teachers who are still developing.” Another teacher predicted a decline in the number of experienced teachers:

Because of all these extra standardized tests and more rigorous evaluations, we're seeing a higher turnover rate of teachers. More [teachers] want to leave the profession within one to five years. You're not getting those ten-plus year teachers any more, hardly ever.

Although the teacher described his or her own reasons for remaining in the profession, citing “the fun, the enjoyment, the love—I love the kids. We love the families. We like the atmosphere,” he or she also acknowledged that increased accountability negatively impacts this view, adding that, “we still enjoy teaching, but it's almost like a burnout.” For one teacher, the negative impact on morale had already prompted some of his or her peers to leave the profession:

Morale is huge. In my head, I am watching at least six really good teachers walk away because this is too much stress for not enough reward anymore. [This] is terrifying to me because I can't imagine doing anything else.

Based on their statements, interviewees had shared concerns that the evaluation system has/will negatively impact teacher retention without perceivably competitive compensation and adequate professional support.

Impact on public perceptions. When asked how teachers' effectiveness classifications/labels, if ever made publically available, would impact parent, student, and/or community perceptions, administrator and teacher interviewees most frequently predicted a further decline in teacher morale and disruptions to the classroom assignment process. An administrator addressed the issue of public access as negatively impacting individual teacher morale and overall community perceptions as follows:

I don't think that [releasing evaluation results] would be a good thing for them to do at all. Because then I just think teachers get beat up enough. I think the school accountability that we have right now is hard enough for some communities.... If you put that on an individual, I think you would have a lot of people leave the profession. I think it would be very hard to retain people.

Another administrator predicted that public access would "be a huge can of worms," adding that "I don't even want to think about it. Hopefully that won't happen. Even the thought of it, you know—somebody gets a rumor out...that teacher's not that good or [another] teacher's really great—now we have to fight with that." A third administrator acknowledged that the school is accountable to parents and the community but cautioned that invalid inferences could be made from widely available evaluation data: "I

understand we're public schools and all of that, but I just think that—I don't think the general public would be able to take that information and digest it the same in the way it's intended necessarily.” Given the complexity of and variation across teacher evaluation systems in Arizona school districts, it would be unreasonable to expect the general public to make valid inferences about teachers as professionals based largely on effectiveness classifications and yet the release of evaluation data in other states foreshadows such use.

Similarly, teacher interviewees also predicted the information would alter parent placement preferences for and administrator perceptions of individual teachers. One noted that, “it would definitely impact the decisions that parents make [in terms of] where they want to place their kids. Even, it might affect the way that principals look at the teachers.” Additionally, the same teacher questioned the inferences that would likely be made based on those classifications asking, “Who’s to tell those parents how those scores were created. For example, for a kindergarten [teacher], again, their scores are directly related to [the] third grade AIMS test. The parents don’t necessarily understand that.”

Another teacher addressed the impact on parent perceptions, especially for a new teacher:

I think that would definitely cause a lot of chaos and drama, and it certainly wouldn’t help any of us grow and become better. If I had somebody come up to me and say, ‘I don’t want my child in your room because you’re this label,’ that would be heartbreaking. I think, especially for those teachers who are new [or] are first year teachers, how horrible that would be to come into this profession, and then that’s your first experience.

In general, teachers were most concerned that parents would not perceive them as effective professionals based on invalid inferences from state-mandated effectiveness

classifications/labels. The following statement, made by an administrator in reference to student development and learning, can also be applied to teacher professional growth:

I know there's a push to run it like a business model.... Say what you will, we're not producing widgets. We're producing people and they are so [much] more variable. They're just not one-size fits all. We're not stamping them out. It's not [like] making a Chevy Malibu. We're making kids and molding kids, [and each is] so completely different. Each one is different. Each teacher is different, and [there's] huge variables in there that I don't know how you'd even measure.

This administrator directly addressed the variability associated with educating individual students within schools as institutions. In the same regard, a teacher evaluation system lacking valid, reliable, and/or fair measures cannot be effectively used as a tool to help individual teachers grow as professionals.

Summary of Results

In this chapter, I discussed the study results for each research question, comparing school administrator and teacher perceptions based on triangulated interview and survey data. Perceptual variations in terms of the evaluation system's purpose, implementation (i.e., intended and actual), measures of effectiveness, and intended and unintended consequences suggested that the standards of effectiveness in the conceptual framework (Cuban, 1998) are useful but applied quite differently by school administrators and teachers. I present warranted assertions for and discuss the applicability of each standard of effectiveness (i.e., purpose, fidelity of implantation, popularity, adaptiveness, and longevity) and overall conclusions in Chapter 6.

CHAPTER 6

Findings and Conclusions

In this chapter, I summarize the study and present findings for each standard of effectiveness (i.e., purpose, fidelity of implementation, popularity, adaptiveness, and longevity) as warranted by the results, both resituated within and supported by the literature (Cuban, 1998). I also discuss overall conclusions and recommend areas for additional research.

Study Summary

Although much research has been conducted on teacher evaluation in recent years, often specifically focused on the use of value-added models to hold teachers accountable for their students' learning, few studies have directly examined variations in the perceptions of stakeholders, namely school administrators and teachers, in a local context regarding the purpose(s) of and implementation processes for a new teacher evaluation system. The ways in which the various stakeholders understand, define, and measure the effectiveness of their teacher evaluation system in practice have not been fully investigated, specifically as situated within a federally-supported, state policy-directed accountability framework.

The purpose of this study was to examine the perceptions of elementary school administrators (i.e., principals and assistant principals) and teachers in a large Arizona school district regarding the use of a new teacher evaluation system, comprised of both professional practice and value-added measures. I sought to better understand how these stakeholders as recipients of, and actors within, a larger, complex policy cycle thus far

measure their system's "value-added." Specifically, I investigated their perceptions of the teacher evaluation system in terms of this conceptual framework.

In order to better understand school administrator and teacher perceptions and to assess the utility of the conceptual framework, I developed a sequential mixed methods research design with two phases of data collection and analyses: stakeholder interviews and large-scale online surveys. Although the surveys were administered to all elementary and secondary school administrators and teachers in the district (as well as other certified staff) as part of a larger evaluation, only elementary administrators and teachers were included in the analyses for this study. Response rates for the administrator and teacher surveys (76.3% and 76.0%, respectively) support claims of representativeness (Nunnally, 1978) as do the results of chi square tests of homogeneity across multiple employment and demographic characteristics.

Using a mixed methods approach, I analyzed all qualitative interview data to inform the development of quantitative survey instruments and then utilized triangulation to seek confirming and disconfirming evidence across the data sources before generating preliminary assertions. I also engaged in various research activities with stakeholder groups to further validate and substantiate my assertions.

In the previous chapter, I organized and presented study results for each construct as aligned to the first three research questions: 1) the system's purpose; 2) intended and actual implementation; and 3) measures of effectiveness including validity, reliability, fairness, and intended/unintended consequences. In response to the fourth research question, I discussed and supported with evidence from multiple data sources perceptual variations among and within stakeholder groups for each construct.

In the next section, I present overall study findings and address the utility of each standard of effectiveness in the conceptual framework (i.e., purpose, fidelity of implementation, popularity, adaptiveness, and longevity) (Cuban, 1998). In addition, I discuss the implications of findings for policymakers, district leadership, and practitioners, specifically arguing that school administrator and teacher perceptions of, and experiences within, the teacher evaluation system exemplify symbolic adaptation to yet another school reform in the implementation phase of the policy cycle (Popkewitz et al., 1982; Tyack & Cuban, 1995). In conclusion, I recommend areas for further research.

Findings and Implications

Much research in recent decades has examined educational policies as micro-level reforms in school contexts and as macro-level processes for reforming schools (David & Cuban, 2010). Undoubtedly, most reformers intend to improve schools for the benefit of all students and sincerely believe that the identified problem, if solved, will accomplish that worthwhile goal (David & Cuban, 2010). Determining the effectiveness of reforms in practice necessitates standards of measurement. Cuban (1998) cited three standards (i.e., purpose, fidelity of implementation, and popularity) as most commonly used by policymakers and others in positions of authority, and argued that practitioners more frequently employ two other standards (i.e., adaptiveness and longevity) when appraising the effectiveness of a reform. Based on the results of this study, I argue that administrators and teachers apportion the standards somewhat differently. In the next section, I present five assertions (one for each standard of effectiveness) and provide evidence for each standard to support the respective assertion.

Purpose. Teacher evaluation systems have been postured in policy talk as possible, logical, and rational means by which teachers can improve their professional practice and ultimately increase student learning (Popkewitz, 1991; Tyack & Cuban, 1995). The largely non-controversial goal of helping teachers grow professionally has received widespread public support (Tyack & Cuban, 1995) and supposedly been reified through federally-supported, state policy-directed teacher evaluation systems. Cuban (1998) suggested that those in positions of authority (e.g., state and national policymakers) determine the primary purpose of a reform and subsequently measure its effectiveness against their desired goals. This argument has important implications for school administrators and teachers as they are the primary recipients of, and actors within, the implementation phase of this particular reform effort (Tyack & Cuban, 1995). Although the district in this study had some autonomy to develop the evaluation system within given parameters, the high-stakes classification labels, arguably the most consequential aspect of the evaluation framework, were prescribed through state legislation.

Assertion 1. School administrators and teachers shared a common belief that teacher evaluation systems in general should be used to improve teachers' professional practice; however, most teachers disagreed that the system in this district was implemented for its stated purpose. This disjunction between the perceptions of school administrators and teachers serves as the keystone for measuring the utility of the other four standards of effectiveness in the conceptual framework for each respective group.

Espousing the view posited as the most logical and rational (Popkewitz, 1991), all school administrators ($n = 29/29$, 100.0%) and the vast majority of teachers ($n = 569/663$,

85.8%) in this district believed that the purpose of evaluating teachers in general should be to improve their professional practice. Some administrator interviewees acknowledged that the system could or arguably should also be used to hold teachers accountable and/or make employment decisions; however, nearly eight out of ten ($n = 23/29$, 79.3%) reiterated on the survey that the primary purpose is to support teacher growth. Substantially fewer teachers ($n = 233/662$, 35.2%) agreed that the evaluation system in place is actually intended for that stated purpose. In fact, nearly half of teachers ($n = 296/662$, 44.7%) indicated that the system was designed and implemented as an accountability mechanism. I argue that school administrators and teachers apply the other standards of effectiveness differently based on their (lack of) confidence in the alignment of purpose, ideally and in reality.

Fidelity of implementation. Cuban (1998) defined fidelity as a measure of “the fit between the initial design, the formal policy, the subsequent program it spawns, and its implementation” (p. 458). When measuring the fidelity of actual to intended implementation, school administrator and teacher perceptions varied considerably depending on the indicator used (e.g., transparency, understanding, completion/participation, utility, etc.). These perceptual variations were particularly relevant when determining the utility of the fidelity standard for each group.

Assertion 2. School administrators generally had a better understanding of evaluation system components and processes, particularly the professional practice measure; were more satisfied with district transparency, communication, and professional development; and overall had greater confidence in their training and ability to conduct evaluations objectively than teachers in this study. Although administrators and teachers

both recognized the importance of these aspect(s), they often disagreed about which, if any, were problematic. Furthermore, even when areas in need of improvement were commonly identified, in some cases, school administrators and teachers did not propose similar solutions.

Administrators generally described the process of developing the system at the district level as transparent; however, perhaps unsurprisingly, teachers frequently assessed transparency (for better or worse) in terms of formal and/or informal communications with school-level rather than district-level administrators. Given this variation, it is important to note that study results suggested communication between school administrators and teachers varied widely across the district. With regards to their understanding of the evaluation system components (e.g., how teachers' professional practice and /or value-added scores are calculated) and/or processes (e.g., clarity of evaluation steps/activities), administrators also reported having greater confidence in the system than teachers. Administrators appraised professional development training and/or resources provided by the district as more helpful than teachers did. Although administrators and teachers reported similar rates of participation in the prescribed evaluation activities, administrators consistently reported higher utility of activities with the exception of the personal self-assessment (completed by each teacher) and end of year conference (during which teachers receive their professional practice score). Administrators also expressed greater confidence that they had been well trained, could evaluate teachers objectively, and had spent enough time in teachers' classrooms.

As mentioned previously, Cuban (1998) defined the fidelity standard as a measure of fit between the initial design, policy, program, and implementation. I argue that the

disparate perceptions of school administrators and teachers with regards to implementation fidelity can be better understood in conjunction with the other standards of effectiveness, particularly the standard of purpose. For the vast majority of administrators, I believe that their assessment of system effectiveness using the fidelity standard reflected their common understanding of the system's purpose—to improve the professional practices of teachers. As such, administrators' strong belief in the alignment between the ideal purpose of evaluation in general and its purpose in this district increased the utility of the fidelity standard for that group. For example, given school administrators' high level of confidence, nearly universal teacher completion of/participation in perceivably clearly defined, useful evaluation activities could indicate a high level of system effectiveness. In addition, administrators generally described the system's development and implementation as transparent, district communication as more than adequate, and professional development as useful preparation to evaluate teachers. Although they acknowledged the significant amount of time required to conduct evaluations, administrators generally did not describe the time currently spent as inadequate. Overall, school administrators agreed that they (to some extent) and certainly teachers would benefit from increased communication and additional professional development training; however, few directly cited these aspects as indicators of poor implementation.

However, given teachers' perceived misalignment of the ideal and actual purposes of the system, completion of/participation in evaluation activities did not seem to be a useful effectiveness measure for them. Unsurprisingly, teachers assessed implementation fidelity in terms of their experiences at their school sites. They seemed to attribute (a lack

of) transparency, communication, and adequate professional development training to formal and informal interactions with their school administrator(s). Even when teachers were individually satisfied with the transparency, communication, and professional development at their school, they either implied, or directly asserted, that these aspects of implementation were likely inconsistent across school sites. In addition, some teachers questioned the ability of school administrators to evaluate them objectively, most frequently citing insufficient time spent by their evaluator in their classroom as exacerbating the problem. Based on study results, teachers placed greater emphasis on these aspects as indicative of implication issues. Given widespread concerns with implementation, I argue that many teachers also applied the other standards (i.e., popularity, adaptiveness, and longevity) to determine system effectiveness (Cuban, 1998).

Popularity. Cuban (1998) cited popularity as one of the primary standards used by policymakers and others in positions of authority to determine the “fashionableness” of reforms among constituents as a prerequisite for their support. This standard is also relevant for school administrators and teachers as recipients of, and actors within, the evaluation system. Popkewitz (1991) noted that power in schooling shapes the ways individuals construct their identities and understand their experiences over time (p. 14). Popkewitz et al. (1982) further described the professional ideology at a school as guiding the behavior of those implementing a reform. School administrators’ and teachers’ professional ideology has been to some extent regulated and reinforced by the state legislature, state department of education, school district, and community (Popkewitz et

al., 1982). However, this power dynamic did not preclude administrators and teachers from measuring system effectiveness themselves based on its popularity.

Assertion 3. School administrators and teachers redefined and utilized the standard of popularity as a measure of system validity, reliability, and fairness. Overall perceptions among many teachers that the system is invalid, unreliable, and/or unfair may be attributable to perceived misalignment of evaluation purposes, ideally and in reality. While some administrators raised similar concerns, they generally assessed the system's popularity in terms of validity, reliability, and fairness quite differently than teachers.

The vast majority of administrators ($n = 27/29$, 93.1% and $n = 24/29$, 82.8%, respectively) indicated that the Danielson FFT is a comprehensive measure of professional practice and/or that a combination of professional practice and student achievement measures is the best indicator of effectiveness. As such, it is important to note that nearly one-fourth of teachers questioned the validity of the Danielson FFT ($n = 171/636$, 24.2%) and/or the use of a combination of measures ($n = 272/636$, 42.7%). In addition, teachers widely expressed concerns about the reliability of ratings and value-added scores in interviews as well as the ability of the system to account for student motivation, attitudes, and engagement ($n = 254/580$, 44.0%) and student background characteristics ($n = 361/579$, 62.3%). They also questioned the fairness of Group A and B classifications ($n = 227/569$, 39.9% and $n = 309/529$, 58.4%, respectively). Again, some administrators also raised these concerns, but they did so to a lesser extent than teachers.

Although 28.6% of administrators ($n = 8/28$) acknowledged that teachers' overall effectiveness classifications/labels may not represent their professional practice, more than four out of ten teachers ($n = 216/512$, 42.2%) directly challenged the use of their

labels to make inferences about the quality of their professional performance.

Unsurprisingly, teachers expressed less confidence in their ability to control, and improve, their future value-added scores and overall effectiveness classifications, and to a lesser extent their professional practice scores, than administrators. I argue that widely-held perceptions among teachers of the system's purpose and popularity as misaligned, invalid, unreliable, and/or unfair prompted them to assess the evaluation system in terms of adaptiveness and longevity (Cuban, 1998).

Adaptiveness. As the “foot-soldiers of every reform aimed at improving student outcomes” (Cuban, 1998, p. 459), teachers, and to a lesser extent school administrators, also measured the effectiveness of the teacher evaluation system using the standard of adaptiveness. Cuban (1998) argued that teachers alter and adapt reforms during implementation, both of which are “healthy signs of inventiveness, active problem solving, and a precondition for determining effectiveness” (p. 460). Based on this argument, teachers' confidence in their own ability to effectively alter and adapt the teacher evaluation system was paramount to successful implementation, and ultimately, the achievement of stated goals.

Assertion 4. School administrators generally agreed that teachers were responsible for and could directly impact their evaluation outcomes. Accordingly, administrators believed that the system has had a generally positive impact on their own and teachers' professional practices as well as student achievement and learning. Teachers overall expressed less confidence in their ability to impact their evaluation outcomes and largely described the system as having had no real or even a negative impact on their professional practice and/or student achievement. As such, many teachers

reportedly have not made significant changes (whether presumably needed or not) in their professional practice as a direct result of their evaluation outcomes.

As previously mentioned, all or nearly all administrators believed that teachers can control, and improve, their future professional practice scores ($n = 27/27$, 100.0%), value-added scores ($n = 26/27$, 96.3%), and/or overall effectiveness classifications ($n = 27/27$, 100.0%). However, teachers were less confident than administrators, specifically with regards to their ability to control, and improve, their own future value-added scores ($n = 278/527$, 52.8%), and/or effectiveness classifications ($n = 351/519$, 67.6%). Given these disparate perceptions, it is unsurprising that 96.2% of school administrators ($n = 25/26$) but less than four out of ten teachers ($n = 218/566$, 38.5%) believed that the system has had a generally positive impact on teachers'/their professional practice. While 84.0% of administrators ($n = 21/25$) described the impact of the system on student achievement and learning as generally positive, less than one-third of teachers ($n = 170/558$, 30.5%) agreed.

Given that many teachers viewed the purpose of the system as misaligned, these results are not surprising. Although the relationships between teachers' perceptions of the system's purpose and its impact on their professional practice and student achievement were not explicitly examined as part of data analyses, it is important to reiterate that at least six out of ten teachers cited the evaluation system as having no real or a generally negative impact in these two areas. Although teachers undoubtedly recognized the high-stakes implications of a poor evaluation score and/or "Ineffective" classification label, these consequences did not appear to have served as the impetus for change in their professional practice (whether presumably needed or not). Based on study results,

teachers' perceived inability to influence their own evaluation outcomes, particularly their value-added scores, left many feeling powerless, essentially serving as passive recipients of their effectiveness classification labels rather than as professional educators actively participating in the evaluation process.

Furthermore, I argue that the current application of and potential for future high-stakes consequences as a result of evaluation outcomes has already spurred some teachers to resist, and in some cases reject altogether, the professional ideology regulated and reinforced by those in positions of authority (e.g., state and national policymakers) (Popkewitz et al., 1982). Study results suggested that most teachers in this district do not characterize formal evaluation processes or related outcomes as foundational to their identity as professional educators who inherently strive for continuous professional improvement. According, I believe that many teachers viewed potential efforts to substantively alter or adapt the system to better suit their needs as ritualistic and/or ceremonial (Popkewitz et al., 1982). Consequently, teachers who neither believed that the teacher evaluation system supported their professional growth nor that it was adaptive to meet their professional needs are most likely to measure system effectiveness using the standard of longevity (Cuban, 1998).

Longevity. Cuban (1998) argued that adaptation by practitioners, essentially the inverse of the fidelity standard, is a necessary prerequisite for a reform to have longevity (p. 460). Reforms that are not adaptive to local conditions are devalued by practitioners. When applied to teacher evaluation frameworks, a perceived lack of adaptiveness erodes the durability of the reform effort in the policy cycle (Cuban, 1998; Tyack & Cuban, 1995).

Tyack and Cuban (1995) suggested that the context in which policy talk occurs within larger cycles changes slowly over time as educational institutions reframe the conversation. Although accountability conversations may change slightly, there are still multiple layers of meaning in measuring the success of reforms in practice (Popkewitz et al., 1982). Popkewitz et al. (1982) argued that, “publically accepted criteria or standards by which people judge success or failure” merely represent a surface layer of meaning (p. 9). School reform is most commonly evaluated in terms of its efficiency in meeting the criteria or standards at the surface. As defined here, the purpose and fidelity of implementation standards applied by school administrators and teachers in this district constituted surface layers.

It is important to note that these surface meanings rarely account for the modification of content and culture that inherently occurs through schooling, the biases and selection that occur in the culture transmission process, and the importance of rituals, ceremonies, and slogans as means of hiding or obscuring the relationship between school practices and social commitment (Popkewitz et al., 1982, p. 11). Rather, the “socially accepted procedures, guidelines, and assumptions” that legitimize reform activities, interactions, and experiences constitute these underlying layers (Popkewitz et al., 1982, p. 9). In this regard, the standards of popularity (as defined and utilized here), adaptiveness, and longevity constituted underlying layers.

When applied in this context, participation in teacher evaluation rituals (e.g., steps in the evaluation process), ceremonies (e.g., the high-stakes rewards and sanctions associated with evaluation outcomes), and slogans (e.g., the classification of teachers according to prescribed effectiveness labels) may have unknowingly or perhaps

unintentionally hid or obscured the lack of meaningful change in teachers' professional practice (Popkewitz et al., 1982). Those in positions of authority who observe reform activities laden with rituals, ceremonies, and slogans may arguably perceive the reform as popular among stakeholders and adaptive to meet their needs. These perceptions support and reinforce conclusions that the reform in practice is aligned to its stated purpose and that it has been implemented with fidelity. Additionally, policymakers may wrongfully conclude that such arguably successful reform efforts have longevity—an assumption that may disregard the perceptions and experiences of practitioners in reality (Cuban, 1998).

Assertion 5. School administrators and teachers substantively differed in their perceptions of system popularity and adaptiveness but recognized similar unintended consequences associated with evaluation processes and outcomes that could threaten system longevity. Despite their shared assessment, school administrators generally emphasized pragmatic concerns about human capital functions (e.g., teacher recruitment, hiring, promotion, and retention). Teachers more often expressed concerns about the affective impacts of evaluation on their professional self-efficacy as well as teacher and/or community morale.

Although cognoscente of threats to the validity, reliability, and fairness of evaluation outcomes, some school administrators expressed their willingness to consider teachers' classification labels when making hiring decisions (Konstantopoulos, 2014). However, more frequently administrators predicted that an increase in teacher turnover would negatively impact staffing, especially cautioning that less experienced teachers with strong potential to be effective or highly effective would be difficult to recruit or

retain without adequate support (Herlihy et al., 2014; Weisberg et al., 2009). Teachers often made anonymous references to peers who had already decided to leave the district or the teaching profession altogether as evidence of this negative impact (whether intended or not) (Herlihy et al., 2014).

In terms of professional self-efficacy, teachers strongly emphasized the negative impact of an “Ineffective” label on an individual teacher. As previously discussed, many teachers disagreed with the vast majority of school administrators, noting that the evaluation processes had little impact on their professional practice thus far. However, teachers generally distinguished between the negligible impact on their practice and the high-stakes implications of using policy-mandated effectiveness classifications to label individual teachers. Although school administrators also cautioned against assigning “Ineffective” labels, especially to those who are new to the profession, most did so in the context of human capital functions (Donaldson, 2011).

School administrators and teachers not only predicted a widespread disruption to school staffing but also a decline in teacher (and if results were publically available, community) morale. School administrators primarily contextualized concerns about lower teacher morale as problematic for retention. In the event that evaluation results were made public, administrators cited negative public perceptions of their school and teachers as highly disruptive to student/teacher assignment processes. In contrast, teachers emphasized the strong likelihood that their school administrators and the general public would make invalid inferences about them as professionals based on their effectiveness classifications (Herlihy et al., 2014; Kane, 2008, 2013; Messick, 1980). Teachers’ generally unfavorable assessment of the system’s long-term viability given

these unintended consequences reflected their widespread frustration and concerns about purpose misalignment and participation in evaluation activities that were perceived as ritualistic, ceremonial, and sloganeered (Cuban, 1998; Popkewitz et al., 1982).

Conclusions

The findings of this study merit close attention from policymakers. The perceptions and lived experiences of school administrators and teachers in this district should be examined in the context of a larger, more complex state and national accountability policy cycle (Amrein-Beardsley & Collins, 2012; Cuban, 1998; David & Cuban, 2010; Tyack & Cuban, 1995). This study contributes to the larger body of research on teacher accountability systems by assessing the utility of the five standards of effectiveness (Cuban, 1998) in the context of current policy trends. At the micro-level, stakeholder confidence (or a lack thereof) in the alignment of this system's purpose, fidelity of implementation, popularity, adaptiveness, and longevity evidenced a schism between evaluators and evaluatees. Even after the vast majority of teachers had completed or participated in time consuming evaluation activities, many still concluded that the system had been largely unsuccessful by these standards. The integration of symbolic adaptation into the conceptual framework also contributes to the literature as an explanation of how stakeholder compliance perhaps unwittingly perpetuates the cycle of ineffective policy talk, action, and implementation (Popkewitz et al., 1982; Tyack & Cuban, 1995).

At the surface, stakeholder participation in evaluation rituals, ceremonies, and slogans served to legitimize teachers and school administrators as professionals and the school district as a credible institution (Popkewitz et al., 1982). However, upon closer

examination, it became clear that many teachers, and perhaps even some school administrators, symbolically adapted their behavior as actors within the evaluation system (Popkewitz et al., 1982). Study results supported the assertion that the system has not uniformly impacted teachers' professional practice in any meaningful way, at least not to the extent intended. Based on this finding, policymakers should not assume that observed behaviors evidence successful implementation and continue to engage in policy talk under the pretense that previous reforms resulted in the desired outcomes. Based on such surface layer evidence, policymakers may erroneously conclude that a reform, when implemented with fidelity based on its stated purpose, yielded meaningful change as intended in varied local contexts. In reality, compliance on the part of implementers in this study may actually serve to perpetuate similar reform efforts based on unsubstantiated claims of effectiveness.

This finding has significant implications not only for district leadership in this study but also for state and national policymakers. There was very little evidence to suggest that district administration in this study developed and implemented the teacher evaluation system without substantive improvement of teacher professional practice as its primary or at least an ancillary goal; however, despite their undoubtedly good intentions, leaders here and in other districts operated within federally-supported, state policy-directed parameters (David & Cuban, 2010). Given this context, it seems that responsibility for the success of teacher evaluation reform efforts should presumably fall upon those who established the policy parameters in the first place. However, I argue that this conclusion is overly simplistic.

Rather than assign blame for arguably unsuccessful reform efforts to either policymakers or implementers, it is critical to understand the standards of effectiveness used by each to measure success (Cuban, 1998). Because standards are defined, applied, and prioritized differently between and among groups (Cuban, 1998), acknowledgement of and open dialogue about the validity, reliability, and fairness of their respective standards should be an integral component of the policy cycle (Tyack & Cuban, 1995). Policymakers should not engage in policy talk alone. Rather, practitioners need to interject their own voices into the conversation, assuming an active role in diagnosing problems and advocating for solutions, that if adapted and implemented in their own districts and schools, could positively impact the schooling experiences of their students (Tyack & Cuban, 1995).

Policymakers also must collaborate with practitioners in the action phase of the policy cycle, especially with school administrators and teachers as they will be expected to assume ownership of the implementation (Tyack & Cuban, 1995). Without the involvement of implementers at each phase, policymakers need not have critical dialogue about the standards of effectiveness they plan to utilize. More often, a narrow set of one-size-fits-all standards (e.g., purpose, fidelity of implementation, and to some extent popularity) becomes the de facto measure of a reform's success (Cuban, 1998)—a judgment of effectiveness that inherently serves to (de)legitimize school administrators and teachers as professionals (Popkewitz et al., 1982). While the chronic failure of policymakers to consider the lived experiences of implementers may be considered a hallmark of education reform (David & Cuban, 2010), school administrators and teachers

share some responsibility to take action if they expect to reframe the conversation about school and teacher accountability.

Recommendations for Further Research

Although this study examined the perceptions of elementary school administrators and teachers in a large Arizona school district, additional aspects of teacher evaluation system development and implementation merit further research. High school administrators and teachers in this district participated in the larger, comprehensive evaluation; however, perceptual variations among and within groups at the high school and elementary levels have not been fully explored. Survey data for other certified staff (e.g., instructional support, counselors, related services, etc.) could offer additional insights. As participants in the evaluation process, their measures of system effectiveness may differ from those of general and special education classroom teachers. A better understanding of their perceptions and experiences would be useful to inform district decision-making.

Further research could broaden the scope of these questions to include other school districts in the state of Arizona and across the nation. Results of this study have greater potential to reframe accountability policy conversations when examined in conjunction with data from other districts and states. Policymakers should have a better understanding of the perceptions of other school administrators and teachers as well given that districts across the state and nation have been tasked with developing and implementing teacher evaluation systems within the same policy-mandated accountability framework.

Comparative studies need to be conducted to inform state and national policymakers, district leaders, practitioners, and the general public before the next cycle of policy talk—a conversation that must be reframed *by* school administrators and teachers if accountability reform efforts are to serve any meaningful purpose in practice (Tyack & Cuban, 1995). As subjects of, and actors within, teacher evaluation policies in context, school administrators and teachers should largely determine and assess the utility of the standards used to measure system effectiveness. Reformers' good intentions are woefully inadequate standards of success in practice.

REFERENCES

- American Educational Research Association. (2000). *AERA position statement: High-stakes testing in preK-12 education*. Washington, DC: American Educational Research Association. Retrieved from <http://www.aera.net/AboutAERA/AERARulesPolicies/AERAPolicyStatements/PositionStatementonHighStakesTesting/tabid/11083/Default.aspx>
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Amrein-Beardsley, A. (2008). Methodological concerns about the Education Value-Added Assessment System. *Educational Researcher*, 37(2), 65-75. doi: 10.3102/0013189X08316420
- Amrein-Beardsley, A. (2014). *Rethinking value-added models in education: Critical perspectives on tests and assessment-based accountability*. New York, NY: Routledge.
- Amrein-Beardsley, A., & Collins, C. (2012). The SAS Education Value-Added Assessment System (SAS® EVAAS®) in the Houston Independent School District (HISD): Intended and unintended consequences. *Education Policy Analysis Archives*, 20. Retrieved from <http://epaa.asu.edu/ojs/article/view/1096>
- Arizona Department of Education. (2011). *Arizona framework for measuring educator effectiveness*. Retrieved from <http://www.azed.gov/teacherprincipal-evaluation/>
- Arizona Revised Statutes §15-203 (A) (38)
- Arizona Revised Statutes §15-537 (I) (3)
- Baker, E. L. (2003). Multiple measures: Toward tiered systems. *Educational Measurement: Issues and Practice*, 22(2), 32–41. doi:10.1111/j.1745-3992.2003.tb00123.x
- Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., Ravtich, D., Rothstein, R., Shavelson, R. J., & Shepard, L. A. (2010). *Problems with the use of student test scores to evaluate teachers*. Washington, DC: Economic Policy Institute. Retrieved from <http://www.epi.org/publication/bp278/>

- Ballou, D. (2012, February 16). Review of “The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood.” Boulder, CO: National Education Policy Center. Retrieved from <http://nepc.colorado.edu/thinktank/review-long-term-impacts>.
- Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., Pianta, R.C., & Qi, Y. (2012). An argument approach to observation protocol validity. *Educational Assessment, 17*, 62-87. doi: 10.1080/10627197.2012.715014
- Berliner, D. C. (2014). Exogenous variables and value-added assessments: A fatal flaw. *Teachers College Record, 116*, 1-31. Retrieved from <http://www.tcrecord.org/Content.asp?ContentId=17293>
- Berliner, D. C., & Biddle, B. J. (1995). *The manufactured crisis: Myths, fraud, and the attack on America’s public schools*. New York, NY: Basic Books.
- Bill & Melinda Gates Foundation. (2010, December). *Learning about teaching: Initial findings from the Measures of Effective Teaching project*. Seattle, WA: Author. Retrieved from http://www.metproject.org/downloads/Preliminary_Findings_Research_Paper.pdf
- Bill & Melinda Gates Foundation. (2012, January). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. Seattle, WA: Author. Retrieved from http://www.metproject.org/downloads/MET_Gathering_Feedback_Practioner_Brief.pdf
- Bill & Melinda Gates Foundation. (2013, January). *Feedback for better teaching: Nine principles for using measures of effective teaching*. Seattle, WA: Author. Retrieved from http://metproject.org/downloads/MET_Feedback%20for%20Better%20Teaching_Principles%20Paper.pdf
- Braun, H. I. (2005). *Using student progress to evaluate teachers: A primer on value-added models*. Princeton, NJ: Educational Testing Service. Retrieved from <http://www.ets.org/Media/Research/pdf/PICVAM.pdf>
- Briggs, D., & Domingue, B. (2011, February). Due diligence and the evaluation of teachers: A review of the value-added analysis underlying the effectiveness rankings of Los Angeles Unified School District teachers by the *Los Angeles Times*. National Education Policy Center. Boulder, CO: University of Colorado at

- Boulder. Retrieved from <http://nepc.colorado.edu/files/NEPC-LAT-VAM-2PP.pdf>
- Burns, R. B., & Mason, D. A. (1995). Organizational constraints on the formation of elementary school classes. *American Journal of Education*, 103(2), 185-212. doi:10.1086/444096
- Capitol Hill Briefing. (2011, September 14). *Getting teacher evaluation right: A challenge for policy makers*. A briefing by E. Haertel, J. Rothstein, A. Amrein-Beardsley, and L. Darling-Hammond. Washington, DC: Dirksen Senate Office Building (research in brief). Retrieved from <http://www.aera.net/Default.aspx?id=12856>
- Chetty, R., Friedman, J., & Rockoff, J. (2014). Discussion of the American Statistical Association's Statement (2014) on using value-added models for educational assessment. Retrieved from http://obs.rc.fas.harvard.edu/chetty/ASA_discussion.pdf
- Childers, K. L. (2012, November). *Holding teachers accountable: Principals' perspectives of the alignment of value-added metrics with evaluations of teacher quality*. Paper presented at the annual meeting of the University Council for Educational Administration, Denver, CO. doi: 10.2139/ssrn.2175931
- Cogan, M. L. (1973). *Clinical supervision*. Boston, MA: Houghton Mifflin.
- Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966). *Equality of educational opportunity*. Washington, DC: United States Department of Health, Education, & Welfare. doi: 10.3886/ICPSR06389.v3
- Corcoran, S. P. (2010). *Can teachers be evaluated by their students' test scores? Should they be? The use of value-added measures of teacher effectiveness in policy and practice*. Providence, RI: Annenberg Institute for School Reform. Retrieved from <http://annenberginstitute.org/publication/can-teachers-be-evaluated-their-students%E2%80%99-test-scores-should-they-be-use-value-added-me>
- Council of Chief State School Officers (CCSSO). (2011, April). *The Interstate Teacher Assessment and Support Consortium (InTASC) Model Core Teaching Standards at a glance*. Washington, DC: Author. Retrieved from <http://www.ccsso.org/Documents/2011/InTASC%202011%20Standards%20At%20A%20Glance.pdf>

- Creative Research Systems. (n.d.). In *Creative Research Systems*. Retrieved from <http://www.surveysystem.com/sscalc.htm>
- Creswell, J. W. (2003). *Research design: Qualitative, quantitative, and mixed methods approaches* (2nd ed.). Thousand Oaks, CA: Sage.
- Creswell, J. W., & Miller, D. L. (2000). Determining validity in qualitative inquiry. *Theory into Practice, 39*(3), 124-130. doi: 10.1207/s15430421tip3903_2
- Creswell, J. W., Plano Clark, V. L., Gutmann, M. L., & Hanson, W. E. (2003). Advanced mixed methods research designs. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social and behavioral research* (pp. 209-240). Thousand Oaks, CA: Sage.
- Cuban, L. (1998). How schools change reforms: Redefining reform success and failure. *Teachers College Record, 99*(3), 453-477. Retrieved from <http://www.tcrecord.org/content.asp?contentid=10273>
- Cuban, L. (2003). *Why is it so hard to get good schools?* New York, NY: Teachers College Press.
- Cuban, L. (2008). *Frogs into princes: Writings on school reform*. New York, NY: Teachers College Press.
- Cubberley, E. P. (1929). *Public school administration: A statement of the fundamental principles underlying the organization and administration of public education* (3rd Ed.). Boston, MA: Houghton Mifflin.
- Czaja, R., & Blair, J. (2005). *Designing surveys: A guide to decisions and procedures* (2nd ed.). Thousand Oaks, CA: Pine Forge Press.
- Daniel, J. (2012). *Sampling essentials: Practical guidelines for making sampling choices*. Thousand Oaks, CA: Sage.
- Danielson, C. (1996). *Enhancing professional practice: A framework for teaching*. Alexandria, VA: Association for Supervision and Curriculum Development (ASCD).
- Danielson, C. (2007). *Enhancing professional practice: A framework for teaching* (2nd ed.). Alexandria, VA: Association for Supervision and Curriculum Development (ASCD).

- Danielson, C. (2013). *The Framework for Teaching Evaluation Instrument*. Princeton, NJ: The Danielson Group. Retrieved from <http://danielsongroup.org/download/?download=448>
- Danielson, C., & McGreal, T. L. (2000). *Teacher evaluation to enhance professional practice*. Alexandria, VA: Association for Supervision and Curriculum Development (ASCD).
- Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E., & Rothstein, J. (2012). Evaluating teacher evaluation. *Phi Delta Kappan*, 93(6), 8-15. doi: 10.1177/003172171209300603
- David, J. L., & Cuban, L. (2010). *Cutting through the hype: The essential guide to school reform* (Rev. ed.). Cambridge, MA: Harvard Education Press.
- Dedoose. (2014). *Web application for managing, analyzing, and presenting qualitative and mixed methods research data* (Version 5.0.11). Los Angeles, CA: SocioCultural Research Consultants, LLC (www.dedoose.com).
- Denzin, N. K. (1978). *The research act: A theoretical orientation to sociological methods* (2nd ed.). New York, NY: McGraw-Hill.
- Dewey, J. (1938). *Experience and education*. New York, NY: Macmillan.
- Dewey, J. (1973). *The philosophy of John Dewey* (J. J. McDermott, Ed.). New York, NY: G. P. Putnam's Sons.
- Donaldson, M. L. (2011, February). *Principals' approaches to developing teacher quality: Constraints and opportunities in hiring, assigning, evaluating, and developing teachers*. Washington, DC: Center for American Progress. Retrieved from http://cdn.americanprogress.org/wp-content/uploads/issues/2011/02/pdf/principal_report.pdf
- Education Commission of the States. (1983). Action for excellence. *Report of the Task Force on Education for Economic Growth*. Retrieved from <http://onlinelibrary.wiley.com/doi/10.1002/pam.4050030217/abstract>.
- Erickson, F. (1986). Qualitative methods in research on teaching. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (pp.119-161). New York, NY: Macmillan.

- Felch, J., Song, J., & Smith, D. (2010, August 14). Who's teaching L.A.'s kids? *Los Angeles Times*. Retrieved from <http://www.latimes.com/local/la-me-teachers-value-20100815-story.html#page=1>
- Fink, A. (1995). *Evaluation for education and psychology*. Thousand Oaks, CA: Sage.
- Geertz, C. (1973). Thick description: Toward an interpretive theory of culture. In C. Geertz (Ed.), *The interpretation of culture*. New York, NY: Basic Books.
- Glaser, B., & Strauss, A. (1967). *The discovery of grounded theory: Strategies for qualitative research*. Chicago, IL: Aldine.
- Glatthorn, A. A. (1984). *Differentiated supervision*. Alexandria, VA: Association for Supervision and Curriculum Development (ASCD). Retrieved from <http://files.eric.ed.gov/fulltext/ED245401.pdf>
- Glickman, C. D., Gordon, S. P., & Ross-Gordon, J. M. (1998). *Supervision of instruction: A developmental approach* (4th ed.). Boston, MA: Allyn and Bacon.
- Goe, L. (2008). *Key issue: Using value-added models to identify and support highly effective teachers*. Washington, DC: National Comprehensive Center for Teacher Quality. Retrieved from <http://www2.tqsource.org/strategies/het/UsingValueAddedModels.pdf>
- Goldhaber, D. (2007). *Everyone's doing it, but what does teacher testing tell us about teacher effectiveness?* [Working paper 9]. Washington, DC: National Center for the Analysis of Longitudinal Data in Education Research, Urban Institute.
- Goldhammer, R. (1969). *Clinical supervision: Special methods for the supervision of teachers*. New York, NY: Holt, Rinehart and Winston, Inc.
- Greene, J. (2007). *Mixed methods in social inquiry*. San Francisco, CA: Jossey-Bass.
- Guba, E. G., & Lincoln, Y. S. (1989). *Fourth generation evaluation*. Newbury Park, CA: Sage.
- Guba, E. G., & Lincoln, Y. S. (1994). Competing paradigms in qualitative research. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 105-117). Thousand Oaks, CA: Sage.
- Hammack, F. M. (1997). Ethical issues in teacher education. *Teachers College Record*, 99(2), 247-265. Retrieved from <http://www.tcrecord.org/Content.asp?ContentID=10256>

- Hanushek, E. A. (1986). The economics of schooling: Production and efficiency in public schools. *Journal of Economic Literature*, 24(3), 1141-1177. Retrieved from <http://hanushek.stanford.edu/sites/default/files/publications/Hanushek%201986%20JEL%2024%283%29.pdf>
- Hanushek, E. A. (1997). Assessing the effects of school resources on student performance: An update. *Educational Evaluation and Policy Analysis*, 19(2), 141-164. doi: 10.3102/01623737019002141
- Harris, D. N. (2009). Would accountability based on teacher value added be smart policy? An evaluation of the statistical properties and policy alternatives. *Education Finance and Policy*, 4, 319-350. doi:10.1162/edfp.2009.4.4.319
- Harris, D. N. (2011). *Value-added measures in education: What every educator needs to know*. Cambridge, MA: Harvard Education Press.
- Harris, D. N., & Sass, T. R. (2009, September). *What makes for a good teacher and who can tell?* [Working paper 30]. Washington, DC: National Center for Analysis of Longitudinal Data in Educational Research, Urban Institute. Retrieved from <http://www.urban.org/uploadedpdf/1001431-what-makes-for-a-good-teacher.pdf>
- Herlihy, C., Karger, E., Pollard, C., Hill, H. C., Kraft, M. A., Williams, M., & Howard, S. (2014). State and local efforts to investigate the validity and reliability of scores from teacher evaluation systems. *Teachers College Record*, 116, 1-28.
- Herringa, S. G., West, B. T., & Berglund, P. A. (2010). *Applied survey data analysis*. Boca Raton, FL: CRC Press. doi: 10.1201/9781420080674
- Hershberg, T., Simon, V. A., & Lea-Kruger, B. (2004). Measuring what matters: How value-added assessment can be used to drive learning gains. *American School Board Journal*, 191(2), 27-31. Retrieved from http://www.cgp.upenn.edu/pdf/measuring_what_matters.pdf
- Hill, H. C., Kapitula, L., & Umland, K. (2011). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal*, 28(3), 794-831. doi: 10.3102/0002831210387916
- Ho, A. D., & Kane, T. J. (2013, January). *The reliability of classroom observations by school personnel*. Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from http://www.metproject.org/downloads/MET_Reliability%20of%20Classroom%20Observations_Research%20Paper.pdf

- Hong, Y. (2010). A comparison among major value-added models: A general model approach. *Dissertation Abstracts International: Section A. Educational Tests and Measurement*, 71(4), 4485.
- Huberman, A. M., & Miles, M. B. (1994). Data management and analysis methods. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative methods* (pp. 428-444). Thousand Oaks, CA: Sage.
- Jacob, A. (2012). Examining the relationship between student achievement and observable teacher characteristics: Implications for school leaders. *International Journal of Educational Leadership Preparation*, 7(3), 1-13. Retrieved from <http://files.eric.ed.gov/fulltext/EJ997469.pdf>
- Jacob, B. A., & Lefgren, L. (2005, June). *Principals as agents: Subjective performance measurement in education*. Cambridge, MA: National Bureau of Economic Research. Retrieved from <http://www.nber.org/papers/w11463>
- Jacob, B. A., & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics*, 26(1), 101-136. doi: 10.1086/522974
- Jarrell, M. G. (2000, November). *Focusing on focus group use in educational research*. Paper presented at the annual meeting of the Mid-South Educational Research Association, Bowling Green, KY. Retrieved from <http://files.eric.ed.gov/fulltext/ED448167.pdf>
- Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher*, 33(7), 14-26. doi: 10.3102/0013189X033007014
- Johnson, R. B., & Turner, L. A. (2003). Data collection strategies in mixed methods research. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social and behavioral research* (pp. 297-319). Thousand Oaks, CA: Sage.
- Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, 38(4), 319-342. doi: 10.1111/j.1745-3984.2001.tb01130.x
- Kane, M. T. (2008). Terminology, emphasis, and utility in validation. *Educational Researcher*, 37(2), 76-82. doi: 10.3102/0013189X08315390
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1-73. doi: 10.1111/jedm.12000

- Koedel, C., & Betts, J. R. (2007). *Re-examining the role of teacher quality in the educational production function* [Working Paper No. 2007-03]. Nashville, TN: National Center of Performance Incentives. Retrieved from http://economics.missouri.edu/working-papers/2007/wp0708_koedel.pdf
- Koedel, C., & Betts, J. R. (2011). Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the Rothstein critique. *Education Finance and Policy* 6(1), 18-42. doi:10.1162/EDFP_a_00027
- Konstantopoulos, S. (2014). Teacher effects, value-added models, and accountability. *Teachers College Record*, 116, 1-21. Retrieved from <http://www.tcrecord.org/content.asp?contentid=17290>
- Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Newbury Park, CA: Sage.
- Little, O., Goe, L., & Bell, C. (2009, April). *A practical guide to evaluating teacher effectiveness*. Washington, DC: National Comprehensive Center for Teacher Quality. Retrieved from <http://www.gtlcenter.org/sites/default/files/docs/practicalGuide.pdf>
- Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V. N., & Martinez, F. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement*, 44(1), 47-67. doi: 10.1111/j.1745-3984.2007.00026.x
- MacQueen, K. M., McLelland, E., Kay, K., & Milstein, B. (1998). Codebook development for team-based qualitative research. *Cultural Anthropology Methods*, 10(2), 31-36. Retrieved from http://www.cdc.gov/hiv/pdf/library_software_answer_codebook.pdf
- Marzano, R. J., Frontier, T., & Livingston, D. (2011). *Effective supervision: Supporting the art and science of teaching*. Alexandria, VA: Association for Supervision and Curriculum Development (ASCD).
- McCaffrey, D. F., Lockwood, J., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics*, 29(1), 67-101. doi:10.3102/10769986029001067
- McCaffrey, D. F., Sass, T. R., Lockwood, J. R., & Mihaly, K. (2009). The intertemporal variability of teacher effect estimates. *Education Finance and Policy*, 4(4), 572-606. doi: 10.1162/edfp.2009.4.4.572

- McGreal, T. L. (1983). *Successful teacher evaluation*. Alexandria, VA: Association for Supervision and Curriculum Development (ASCD).
- Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist*, *30*, 955-966. doi: 10.1037/0003-066X.30.10.955
- Messick, S. (1980). Test validity and the ethics of assessment. *American psychologist*, *35*(11), 1012-1027. doi: 10.1037/0003-066X.35.11.1012
- Miles, M. B., & Huberman, A. M. (1984). Drawing valid meaning from qualitative data: Toward a shared craft. *Educational Researcher*, *13*(5), 20-30. doi: 10.3102/0013189X013005020
- Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis* (2nd ed.). Thousand Oaks, CA: Sage Publications.
- Monk, D. H. (1987). Assigning elementary pupils to their teachers. *Elementary School Journal*, *88*(2), 167-187. doi:10.1086/461531
- National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform*. Washington, DC: United States Government Printing Office.
- Newton, X. A., Darling-Hammond, L., Haertel, E., & Thomas, E. (2010). Value-added modeling of teacher effectiveness: An exploration of stability across models and contexts. *Educational Policy Analysis Archives*, *18*(23). Retrieved from <http://epaa.asu.edu/ojs/article/view/180>
- Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw Hill.
- Papay, J. P. (2010). Different tests, different answers: The stability of teacher value-added estimates across outcome measures. *American Educational Research Journal*, *48*(1), 163-193. doi: 10.3102/00002831210362589
- Patton, M. Q. (1990). *Qualitative research and evaluation methods* (3rd ed.). Thousand Oaks, CA: Sage.
- Paufler, N. A., & Amrein-Beardsley, A. (2014). The random assignment of students into elementary classrooms: Implications for value-added analyses and interpretations. *American Educational Research Journal*, *51*(2), 328-362. doi: 10.3102/0002831213508299

- Player, D. (2010). Nonmonetary compensation in the public teacher labor market. *Education Finance and Policy*, 5(1), 82-103. doi:10.1162/edfp.2009.5.1.5105
- Plummer, K. (2001). The moral and human face of life stories: Reflexivity, power, and ethics. In K. Plummer (Ed.), *Documents of life 2: An invitation to critical humanism* (pp. 204-231). Thousand Oaks, CA: Sage.
- Popham, W. J. (1988). *Educational evaluation* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.
- Popkewitz, T. S. (1991). *A political sociology of educational reform: Power/knowledge in teaching, teacher education, and research*. New York, NY: Teachers College Press.
- Popkewitz, T. S. (2008). *Cosmopolitanism and the age of school reform: Science, education, and making society by making the child*. New York, NY: Routledge.
- Popkewitz, T. S., Tabachnick, B. R., & Wehlage, G. (1982). *The myth of educational reform: A study of school responses to a program of change*. Madison, WI: University of Wisconsin Press.
- Praisner, C. (2003). Attitudes of elementary school principals toward the inclusion of students with disabilities. *Exceptional Children*, 69(2), 135-145. doi: 10.1177/001440290306900201
- Rossi, P. H., Lipsey, M. W., & Freeman, H. E. (2004). *Evaluation: A systematic approach* (7th ed.). Thousand Oaks, CA: Sage.
- Rothstein, J. (2009). *Student sorting and bias in value-added estimation: Selection on observables and unobservables*. Cambridge, MA: The National Bureau of Economic Research. Retrieved from <http://www.nber.org/papers/w14607>
- Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *Quarterly Journal of Economics*, 125(1). doi:10.1162/qjec.2010.125.1.175
- Sanders, W. L. (2006, October). *Comparisons among various educational assessment value-added models*. Paper presented at the Power of Two – National Value-Added Conference, Columbus, OH. Retrieved from <http://www.sas.com/resources/asset/vaconferencepaper.pdf>

- Sanders, W. L., & 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. doi: 10.1023/A:1008067210518
- Sanders, W. L., & Rivers, J. C. (1996, November). *Cumulative and residual effects of teachers on future student academic achievement*. Knoxville, TN: University of Tennessee Value-Added Research and Assessment Center. Retrieved from <http://heartland.org/policy-documents/cumulative-and-residual-effects-teachers-future-student-academic-achievement>
- Sanders, W. L., Wright, S. P., Rivers, J. C., & Leandro, J. G. (2009). *A response to criticisms of SAS EVAAS*. Cary, NC: SAS Institute Inc. Retrieved from www.sas.com/resources/asset/Response_to_Criticisms_of_SAS_EVAAS_11-13-09.pdf
- Sartain, L., Stoelinga, S. R., & Brown, E. R. (2011). *Rethinking teacher evaluation in Chicago: Lessons learned from classroom observations, principal-teacher conferences, and district implementation*. Chicago, IL: University of Chicago Urban Education Institute, Consortium on Chicago School Research. Retrieved from <http://ccsr.uchicago.edu/sites/default/files/publications/Teacher%20Eval%20Report%20FINAL.pdf>
- Schaeffer, B. (2004, December). Districts pilot value-added assessment. *The School Administrator*. Retrieved from <http://www.aasa.org/publications/saarticledetail.cfm?ItemNumber=1066>
- Scherrer, J. (2011). Measuring teaching using value-added modeling: The imperfect panacea. *NASSP Bulletin*, 95(2), 122-140. doi: 10.1177/0192636511410052
- Schofield, J. W. (1993). Increasing the generalizability of qualitative research. In M. Hammersley (Ed.), *Social research: Philosophy, politics, and practice* (pp. 200-225). Thousand Oaks, CA: Sage.
- Schwant, T. A. (1997). *Qualitative inquiry: A dictionary of terms*. Thousand Oaks, CA: Sage. doi: 10.1177/107780049700300101
- Shavelson, R. J., Webb, N. M., & Burstein, L. (1986). Measurement of teaching. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 50-91). New York, NY: McMillan.

- Sloat, E. (2014). *Examining the validity of a state policy-directed framework for evaluating teacher instructional quality: Informing policy, impacting practice*. (In-progress Doctoral dissertation. Mary Lou Fulton Teachers College, Arizona State University, Phoenix, AZ.
- Smith, M. L. (1997). Mixing and matching: Methods and models. *New Directions for Evaluation*, 74, 73-85. doi: 10.1002/ev.1073
- Spradley, J. (1979). *The ethnographic interview*. San Diego, CA: Harcourt, Brace, and Janovich.
- Stake, R. E. (1978). The case study method in social inquiry. *Educational Researcher*, 7(2), 5-8. doi: 10.3102/0013189X007002005
- Strauss, A. L., & Corbin, J. (1995). *Basics of qualitative research: Grounded theory procedures and techniques*. Newbury Park, CA: Sage.
- Strauss, A. L., & Corbin, J. (1998). *Basics of qualitative research: Techniques and procedures for developing grounded theory* (2nd ed.). Thousand Oaks, CA: Sage.
- SurveyMonkey Inc. (2014). In *surveymonkey.com*. Retrieved from <http://www.surveymonkey.com>
- Swearingen, M. (1946). Looking at supervision. *Educational leadership*, 3(4), 146-151. Retrieved from http://www.ascd.org/ASCD/pdf/journals/ed_lead/el_194601_swearingen.pdf
- Tashakkori, A., & Teddlie, C. (2003). *Handbook of mixed methods in social and behavioral research*. Thousand Oaks, CA: Sage.
- Taylor, E. S., & Tyler, J. H. (2012). The effect of evaluation on teacher performance. *American Economic Review*, 102(7), 3628-3651. doi: 10.1257/aer.102.7.3628
- Taylor, F. W. (1911/1998). *The principles of scientific management*. Norcross, GA: Engineering & Management Press.
- Teddlie, C., & Tashakkori, A. (2006). A general typology of research designs featuring mixed methods. *Research in the schools*, 13(1), 12-28. Retrieved from http://www.msra.org/Rits_131/Teddlie_Tashakkori_131.pdf
- Toch, T., & Rothman, R. (2008). *Rush to judgment: Teacher evaluation in public education*. Washington, DC: Education Sector. Retrieved from

- <http://www.educationsector.org/publications/rush-judgment-teacher-evaluation-public-education>
- Tracy, S. J. (1995). How historical concepts of supervision relate to supervisory practices today. *The Clearing House*, 68(5), 320-325.
- Tucker, P. D., & Stronge, J. H. (2005). *Linking teacher evaluation and student learning*. Alexandria, VA: Association for Supervision and Curriculum Development (ASCD).
- Tyack, D., & Cuban, L. (1995). *Tinkering toward utopia: A century of public school reform*. Cambridge, MA: Harvard University Press.
- Tyler, J. H., Taylor, E. S., Kane, T. J., & Wooten, A. L. (2010). Using student performance data to identify effective classroom practices. *The American Economic Review*, 100(2), 256-260. doi: 10.1257/aer.100.2.256
- Tynan, A. C., & Drayton, J. L. (1988). Conducting focus group – A guide for first-time users. In T. J. Hayes & C. B. Tatham (Eds.), (1989) *Focus group interviews: A reader* (2nd ed., pp. 5-9). Chicago, IL: American Marketing Association.
- United States Department of Education. (2009). *Race to the top program: Executive summary*. Retrieved from <http://www2.ed.gov/programs/racetothetop/executive-summary.pdf>
- United States Department of Education. (2010). *Teacher incentive fund*. Retrieved from <http://www2.ed.gov/programs/teacherincentive/index.html>
- Vaughn, S., Schumm, J. S., & Sinagub, J. (1996). *Focus group interviews in education and psychology*. Thousand Oaks, CA: Sage.
- Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The Widget Effect: Our national failure to acknowledge and act of differences in teacher effectiveness* (2nd ed.). Brooklyn, NY: The New Teacher Project (TNTP). Retrieved from <http://tntp.org/ideas-and-innovations/view/the-widget-effect>
- Wetzel, W. (1929). Scientific supervision and curriculum building. *The School Review*, 37(2), 179-192. doi: 10.1086/438807
- Wise, A. E., Darling-Hammond, L., McLaughlin, M. W., & Bernstein, H. T. (1985). Teacher evaluation: A study of effective practices [Special issue]. *The Elementary School Journal*, 86(1), 60-121. doi: 10.1086/461437

APPENDIX A

ASSOCIATION FOR SUPERVISION AND CURRICULUM DEVELOPMENT:

COPYRIGHT PERMISSION FOR USE

From: **Permissions** <permissions@ascd.org>
Date: Wed, Nov 12, 2014 at 11:45 AM
Subject: RE: Copyright Permission Request (Thread: 1299634)
To: noelle.paufler@asu.edu

In response to your request below, please consider this permission to use the excerpt(s) from the referenced publication for your personal research purposes. Should you include excerpts or cite content in a paper or some other report form, please credit the source accordingly. If your research results in use of our content in a product or publication for commercial release, please contact me again to secure further rights to do so.

Thank you for your interest in ASCD and good luck with your dissertation.

Sincerely yours,

KATY WOGEC • Sr. Paralegal

1703 N. Beauregard Street • Alexandria, VA 22311-1714

P [703-575-5749](tel:703-575-5749) · F [703-575-3926](tel:703-575-3926) · www.ascd.org · www.wholechildeducation.org



From: Noelle Paufler [mailto:noelle.paufler@asu.edu]
Sent: Tuesday, November 11, 2014 8:46 PM
To: permissions@ascd.org
Subject: Copyright Permission Request (Thread: 1299634)

To Whom It May Concern,

I am writing to request permission to paraphrase materials from the following source in my doctoral dissertation:

Marzano, R. J., Frontier, T., & Livingston, D. (2011). *Effective Supervision: Supporting the Art and Science of Teaching*. Alexandria, VA: Association for Supervision & Curriculum Development.

Specifically, I would like to include paraphrased content (approximately 1,500 words) with appropriate citation of the above-referenced and original sources for:

Chapter 2

The Early Days of Supervision and Evaluation (p. 12-13)

The Period of Scientific Management (p. 14-15)

Post-World War II (p. 16-17)

The Era of Clinical Supervision (p. 17-19)

The Era of Developmental/Reflection Models (21-22)

Please feel free to contact me with any questions or concerns.

Thank you for your time and consideration.

Sincerely,

Noelle Paufler
PhD Candidate
Educational Policy and Evaluation
Mary Lou Fulton Teachers College
Arizona State University

APPENDIX B

ASSOCIATION FOR SUPERVISION AND CURRICULUM DEVELOPMENT:

COPYRIGHT PERMISSION TO REPRINT

From: **Permissions** <permissions@ascd.org>
Date: Tue, Sep 30, 2014 at 9:28 AM
Subject: FW: Request for Permission (Thread: 1291567)
To: noelle.paufler@asu.edu

In response to your request below, please consider this permission to adapt the excerpt from the referenced publication for your personal research purposes. Should you include excerpts or cite content in a paper or some other report form, please credit the source accordingly, as you did in the attachment. If your research results in use of our content in a product or publication for commercial release, please contact me again to secure further rights to do so.

Thank you for your interest in ASCD and good luck with your dissertation.

Sincerely yours,

KATY WOGEC • Sr. Paralegal

1703 N. Beauregard Street • Alexandria, VA 22311-1714

P [703-575-5749](tel:703-575-5749) · F [703-575-3926](tel:703-575-3926) · www.ascd.org · www.wholechildeducation.org



From: Noelle Paufler [mailto:noelle.paufler@asu.edu]
Sent: Sunday, September 28, 2014 5:11 PM
To: permissions@ascd.org
Subject: Request for Permission (Thread: 1291567)

To Whom It May Concern,

I am a doctoral candidate in the Mary Lou Fulton Teachers College at Arizona State University, and for my dissertation, I am examining the perceptions of school administrators and teachers regarding the design and implementation of a new teacher evaluation system in a large Arizona school district.

As the evaluation system includes the Danielson Framework for Teaching as a measure of professional practice, I am writing to request permission to use an adapted figure (with appropriate citation) from *Enhancing Professional Practice: A Framework for Teaching* (1996), in my dissertation and any subsequent related publications. I have adapted the figure, titled "A Blueprint for Teacher Evaluation: Components of Professional Practice," by updating the language for the following components to reflect the 2007 edition of the Framework for Teaching:

- 1c: Selecting Instructional Goals to Setting Instructional Outcomes
- 1f: Assessing Student Learning to Designing Student Assessments
- 3a: Communicating Clearly and Accurately to Communicating with Students
- 3d: Providing Feedback to Students to Using Assessment in Instruction
- 4d: Contributing to the School and District to Participating in the Professional Community

I have attached the adapted figure for your review.

Please feel free to contact me with any questions or concerns. I sincerely appreciate your time and consideration.

Sincerely,

Noelle Paufler
PhD Candidate
Educational Policy and Evaluation
Mary Lou Fulton Teachers College
Arizona State University

APPENDIX C

MID-SOUTH EDUCATIONAL RESEARCH ASSOCIATION:

COPYRIGHT PERMISSION TO REPRINT



RESEARCH IN THE SCHOOLS

An internationally refereed journal sponsored by
the Mid-South Educational Research Association
and Sam Houston State University

September 25, 2014

Anthony J. Onwuegbuzie
Sam Houston State University
Co-Editor

John R. Slate
Sam Houston State University
Co-Editor

Larry G. Daniel
University of North Florida
Consulting Editor

Rebecca K. Frels
Sam Houston State University
Production Editor/Editorial
Assistant

Gail H. Hughes
University of Arkansas, Little Rock
Copy Editor

Noelle Paufler
PhD Candidate
Educational Policy and Evaluation
Mary Lou Fulton Teachers College
Arizona State University

RE: Copyright Permission to Reproduce Figure in Dissertation

Dear Jai Seaman:

I am in receipt of your request for copyright permission to reproduce Figure 5 on page 22 of the following article:

Teddle, C., & Tashakkori, A. (2006). A general typology of research designs featuring mixed methods. *Research in the Schools*, 13(1), 12-28.

Research in the Schools (RITS) is owned and copyrighted by the Mid-South Educational Research Association (MSERA). As Co-Editors of RITS, we are given authority to make decisions for MSERA regarding issues of copyright permissions for RITS.

We grant you copyright permission to reproduce this figure in your dissertation. Specifically, we grant you non-exclusive world rights in all media and all languages with no obligation to pay MSERA royalties, providing you and your publisher/copyright holder agree to the following:

- Appropriate attribution must be given to the author, the journal, and the copyright holder.
- A statement similar to the following must accompany the article:
"Reprinted from *Research in the Schools*, © 2006 by the Mid-South Educational Research Association, Nashville, TN. Reprinted with permission of the original copyright holder."

Sincerely,

Anthony J. Onwuegbuzie, Ph.D., Co-Editor
Research in the Schools

cc: John R. Slate, Ph.D., Co-Editor
Eunjin Hwang, Editorial Assistant/Production Editor

Editorial Office Address:
Anthony J. Onwuegbuzie
College of Education
Sam Houston State University
Room 325 Box 2119
Huntsville, Texas 77341-2119
Phone: (936) 294-4509 FAX: (936) 294-3886
email: tonyonwuegbuzie@aol.com

APPENDIX D

ARIZONA STATE UNIVERSITY INSTITUTIONAL REVIEW BOARD APPROVAL

EXEMPTION GRANTED

Audrey Beardsley
 Division of Educational Leadership and Innovation - West
 602/543-6374
 audrey.beardsley@asu.edu

Dear Audrey Beardsley:

On 1/10/2014 the ASU IRB reviewed the following protocol:

Type of Review:	Initial Study
Title:	Investigating a Teacher Evaluation System: Principal and Teacher Perceptions of the System's Standards of Effectiveness
Investigator:	Audrey Beardsley
IRB ID:	STUDY00000467
Funding:	None
Grant Title:	None
Grant ID:	None
Documents Reviewed:	<ul style="list-style-type: none"> • Participant Informed Consent Form.pdf, Category: Consent Form; • Investigating a Teacher Evaluation System, Category: IRB Protocol; • Interview Protocol - School Administrator.pdf, Category: Measures (Survey questions/Interview questions /interview guides/focus group questions); • Interview Protocol - Teacher.pdf, Category: Measures (Survey questions/Interview questions /interview guides/focus group questions); • Teacher Evaluation System Survey Draft.pdf, Category: Measures (Survey questions/Interview questions /interview guides/focus group questions); • Focus Group Protocol Draft.pdf, Category: Measures (Survey questions/Interview questions /interview guides/focus group questions); • Interview Protocol - District Administrator.pdf, Category: Measures (Survey questions/Interview

	questions /interview guides/focus group questions); • District Permission to Conduct Research.pdf, Category: Off-site authorizations (school permission, other IRB approvals, Tribal permission etc); • Interview-Focus Group Participation Letter.docx, Category: Recruitment Materials; • Survey Participation Letter.pdf, Category: Recruitment Materials; • Survey Participation Reminder Letter.pdf, Category: Recruitment Materials;
--	---

The IRB determined that the protocol is considered exempt pursuant to Federal Regulations 45CFR46 (1) Educational settings, (2) Tests, surveys, interviews, or observation on 1/10/2014.

In conducting this protocol you are required to follow the requirements listed in the INVESTIGATOR MANUAL (HRP-103).

Sincerely,

IRB Administrator

cc: Noelle Paufler
Noelle Paufler

APPENDIX E

DISTRICT RESEARCH APPROVAL

Amended: November 3, 2014
Approved: January 9, 2014

RE: Permission to Conduct Research [REDACTED]

Dissertation Topic: Investigating a Teacher Evaluation System: School Administrator and Teacher Perceptions of the System's Standards of Effectiveness

Dear [REDACTED]

As you may know, I am currently a doctoral candidate in the Educational Policy and Evaluation (EPE) program within the Mary Lou Fulton Teachers College at Arizona State University. This spring, I am entering into the dissertation phase of my studies and would like your permission to conduct research within [REDACTED] ASU's EPE program emphasizes the need to conduct scholarly research examining the development and implementation of educational policies in organizational settings, ideally to inform local policy decisions and improve implementation processes. [REDACTED]

The purpose of my study is to examine the perceptions of school administrators (principals and assistant principals) and teachers (and other certificated staff) with regards to the new state policy-directed teacher evaluation system as implemented within the District. Specifically, I hope to gain further, in-depth understandings about the ways that these groups define and measure the effectiveness of the teacher evaluation system (i.e., with regards to the System's purpose, fidelity of implementation, popularity, adaptiveness, and longevity). I hope that the results from my research will not only help inform and improve the development and implementation of [REDACTED] system but also contribute to the broader state policy debate on accountability systems.

In order to more closely examine these perceptions, this study would address the following research questions:

1. What do principals, assistant principals, and teachers perceive as the purpose of the teacher evaluation system?
2. To what extent do principals, assistant principals, and teachers perceive the system as having been implemented with fidelity?
 - a. What do principals, assistant principals, and teachers perceive as the consequences (intended and unintended) of implementation?
3. To what extent is the system popular among principals, assistant principals, and teachers as recipients of and actors within this policy-directed reform?
4. To what extent do principals, assistant principals, and teachers perceive the system as adaptive in terms of professional practice?
5. To what extent do principals, assistant principals, and teachers perceive the system as having the longevity necessary to be effective in a cyclical policy environment?

Research Activity: To conduct my research, I am requesting permission to do the following activities:

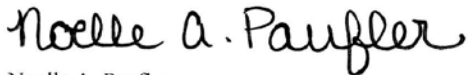
1. Conduct one-on-one interviews with a random sample of selected principals, assistant principals, and teachers as well as district administrators who are connected with the teacher evaluation system development and implementation processes.
2. Administer online surveys to all school administrators and teachers (and other certificated staff) with questions related to teacher evaluation.
3. Conduct focus groups with a sample of selected teachers to discuss aggregate survey results.

Voluntary Participation: In accordance with established ethical guidelines for conducting human subjects research, staff participation in research activities will be completely voluntary based on informed consent. To ensure this, I will provide each selected participant with written communication stating the purpose and intent of the study, assurances of confidentiality, and use of the data in the research process.

Confidentiality: All of the data collected will be protected and confidential including demographic data. No individual participant's name will be released in any form including in the publication of my dissertation, conference presentations, reports, or any subsequent publication. Confidentiality will be maintained by utilizing codebooks and reference identifiers which will be stored in separate electronic and physical locations from the raw data. No student information will be collected or examined for the purposes of this study. In addition, no reference to [REDACTED] will be made in any publication or presentation, and [REDACTED]

I hope you will consider my request to conduct research within the [REDACTED] would be happy to answer any questions you might have and provide more detailed documentation of the research design, research questions, instruments, and analytic plan for your review. I can also provide a copy of ASU's Human Subjects Research IRB application and approval.

Thank you for your time and consideration.



Noelle A. Paufler
PhD Candidate
Educational Policy and Evaluation
Mary Lou Fulton Teachers College
Arizona State University

Approval to Conduct Research within the District:

My signature below represents approval to conduct research within the [REDACTED] within the framework outlined in this communication. It is understood that no personally identifiable information will be released in any form. In addition, no connection of the data, results, or interpretive findings will be made to the [REDACTED]



APPENDIX F
INTERVIEW PARTICIPATION LETTER

Dear _____:

The [name of district removed] is conducting an on-going evaluation of the district's Teacher Evaluation System. As part of the evaluation process, interviews are being conducted with principals, assistant principals, district administrators, and a group of randomly selected classroom teachers. The purpose is to obtain feedback from all stakeholder groups to help decision makers improve the implementation and effectiveness of the program.

To this end, I would like to conduct an interview with you regarding your perceptions of the current evaluation process. The interview would last approximately 45 minutes and be scheduled at your convenience so as not to interfere with classroom instructional time. With your approval, I would like to audio-record the interview to allow for transcription and accurate data analysis.

Your contribution will be combined with feedback received from many other stakeholders. No individually identifiable information will be released in any form.

If you agree to participate, please respond to this email so that I may schedule an interview sometime in the next few weeks.

Voluntary Participation:

Your participation in this research process is completely voluntary. Your principal/supervisor is NOT being informed that (1) you have been randomly selected, or (2) that you have agreed/declined to participate. The [name of department removed] will only inform school administrators that individuals throughout the district have been randomly selected to receive an interview invitation.

Confidentiality:

All information collected is confidential. In this regard, I have attached a copy of the [name of district removed] Confidentiality Statement for your review. This form is used by the [name of department removed] for all research activities. I have also attached a copy of an Arizona State University (ASU) Participant Informed Consent Form requesting your approval to use the information in an academic dissertation, presentations and publications (ASU Institutional Review Board ID: STUDY00000467).

Each document stipulates that no personally identifiable information will be disclosed in any form including disclosure to district or school administrators, supervisors, or colleagues. The only individuals that will have access to the data will be members of the [name of department removed] responsible for collecting, processing, and summarizing the information in a district level evaluation report.

I hope you will consider contributing your perspective on this important topic. Please respond so that we might schedule a convenient time to meet.

I would be happy to answer any questions you might have or provide additional details regarding the research process.

Thank you for your time and consideration.

Sincerely,

Noelle A. Paufler

APPENDIX G

PARTICIPANT INFORMED CONSENT FORM

Dear Participant:

My name is Noelle Paufler, and I am a doctoral candidate working under the advisement of Associate Professor Audrey Amrein-Beardsley in the Mary Lou Fulton Teachers College at Arizona State University. For my doctoral dissertation, I am conducting a research study to examine the perceptions of elementary principals, assistant principals, and teachers with regards to the new state policy-directed teacher evaluation system as implemented within the District. I would like to personally invite you to participate in this study.

Purpose: The purpose of my study is to help build an understanding of the perceptions of principals, assistant principals, and teachers with regards to various aspects of the purpose, implementation, and effectiveness of the [name of district removed] Teacher Evaluation System. Findings will directly inform local policy decisions and help improve [name of district removed] Teacher Evaluation System implementation processes.

Participation: If you agree to participate, I would like to conduct an interview lasting approximately 45 minutes. With your permission, I would also like to audio record the interview. However, I will not do so without your explicit prior consent. Even if you have provided consent, you may change your mind even after the interview is in progress. I would be happy to provide you with advance copies of the interview questions.

Voluntary Participation: Please note that your participation is completely voluntary. If you choose not to participate or to withdraw from the study at any time, for any reason, there will be no penalty. You may also elect not to answer any specific questions. At any time, you may direct me not to utilize any/all of the information that you have provided. There are no foreseeable risks or discomforts to your participation.

Confidentiality: All of the data/information collected in this study is strictly confidential. I will not disclose your participation or any information that you may provide. All data/information will be combined with responses from other participants and analyzed/reported in aggregate form. Your name, position title, location, or other identifiable references will not be released or published in any form. All electronic data and written notes or other documents will be maintained by this researcher in a secure location.

Dissemination of Information: The results of this study will be published as a doctoral dissertation. Copies of the dissertation will be provided to district policymakers and will also be available to all stakeholders participating in the study. In addition, the data/information collected may be utilized in reports, presentations, or publications; however, your name will not be disclosed.

If you have any questions concerning this research study and/or your contribution, please feel free to contact me directly at noelle.paufler@asu.edu.

If you have any questions about your rights as a subject/participant in this research, or if you feel you have been placed at risk, you can contact the following:

- Principal Investigator: Dr. Audrey Amrein-Beardsley, Associate Professor, Arizona State University at audrey.beardsley@asu.edu or 602-561-4731.
- Chair of the Human Subjects Institutional Review Board, Arizona State University Office of Research Integrity and Assurance, at 480-965-6788. (IRB ID: STUDY00000467)

Thank you for your time and consideration.

Sincerely,

Noelle A. Paufler
 Ph.D. Candidate
 Educational Policy and Evaluation
 Mary Lou Fulton Teachers College
 Arizona State University

Your signature below indicates that you consent to participate in the above study.

Signature	Printed Name	Date

RESEARCHER’S STATEMENT

“I certify that I have explained to the above individual the nature, purpose, potential benefits, and possible risks associated with participation in this research study. In addition, I have answered all questions and/or concerns raised and have witnessed the above signature. These elements of Informed Consent conform to the Assurance given by Arizona State University to the Office for Human Research Protections to protect the rights of human subjects. Finally, I have provided the subject/participant a copy of this signed consent document.”

Signature	Printed Name	Date

APPENDIX H
INTERVIEW REMINDER LETTER

Dear _____:

I wanted to follow-up with you regarding my request for an interview on teacher evaluation to see whether you would like to participate. I am very interested in your perspective and would greatly appreciate your feedback.

Please respond to this email so that I may schedule an interview at your convenience.

Please also feel free to contact me with any questions or concerns.

Thank you again for your time and consideration.

Sincerely,

Noelle A. Paufler

APPENDIX I

SCHOOL ADMINISTRATOR INTERVIEW PROTOCOL

Date & Location: _____
Interviewee Identification: _____
Position: School Administrator
Time: Approximately 45 minutes per session
Location: Face-to-face; School, District Office, etc.
Method of Data Collection: Audio-recording, interview notes, artifacts

Interview Session Activities:

- Purpose of Interview (i.e., background, purpose, expected time frame)
- Review of [name of district removed] Confidentiality Statement and Dissertation Participation Informed Consent Form (i.e., obtain signature)
- Interview Activity
- Discussion of Data Analyses (e.g., transcription, inclusion in [name of district removed] Evaluation and doctoral dissertations, review and coding processes)

Interview Questions: (In your opinion...)

1. Purpose of the [name of district removed] Teacher Evaluation System
 - SYSTEM DESIGN AND IMPLEMENTATION (*Organizational*)
 - What is the purpose of the teacher evaluation system?
 - Do you believe school administrators share a common understanding of the purpose of the system? Why or why not?
 - To what extent have the processes for designing and implementing the evaluation system at the district-level been transparent?
 - How much input do you feel school administrators have had in the design and implementation processes?
 - TEACHER EVALUATION PROCESS (*Individual*)
 - Do you believe that the steps of the teacher evaluation process have been clearly defined?
 - Do you feel well prepared to evaluate teachers?
 - What part of the teacher evaluation process do you value most? Why?
2. Measuring Teacher Quality
 - Do you believe the Danielson FFT measures the most important aspects of teacher quality? Are there any domains or components you think are missing?
 - Do you believe there is consistency among evaluators across the district? Why or why not?
 - Is the teacher evaluation process applied fairly to all teachers?
 - Do you believe the value-added model is a good measure of teacher quality? Why or why not?
 - If teachers' professional practice scores and value-added scores are not aligned, in which would you place more confidence? Why?

- In general, do you believe teachers at your school (based on the pilot year data) received final classifications/labels that reflect their level of effectiveness?

3. Impact on Professional Practice

- How has participation in the teacher evaluation process impacted you professionally and personally?
- Has it changed your professional practice? If so, in what ways?
- Has the evaluation process impacted the professional practice of teachers at your school thus far? If so, in what ways?
- What impact, if any, do you think teachers' final classifications/labels will have on teacher hiring and retention at your school?
- What impact, if any, do you think teachers' final classifications/labels will have on the perceptions of parents, students, and others in the community?

4. Improving Implementation

- In what ways, if any, can teacher evaluation be improved?
- How can the teacher evaluation system as a whole be improved?
- How can the evaluation process be improved?
- What additional training, if any, would be helpful for you in terms of the Danielson FFT rubric or the overall teacher evaluation process?

5. Additional Comments

- Is there anything you would like to add?

APPENDIX J

TEACHER INTERVIEW PROTOCOL

Date & Location: _____
Interviewee Identification: _____
Position: Teacher
Time: Approximately 45 minutes per session
Location: Face-to-face; School, District Office, etc.
Method of Data Collection: Audio-recording, interview notes, artifacts

Interview Session Activities:

- Purpose of Interview (i.e., background, purpose, expected time frame)
- Review of [name of district removed] Confidentiality Statement and Dissertation Participation Informed Consent Form (i.e., obtain participant's signature)
- Interview Activity
- Discussion of Data Analyses (e.g., transcription, inclusion in [name of district removed] Evaluation and doctoral dissertations, review and coding processes)
- Request for Transcript Review (e.g., member-checking processes – interviewee verification, modifications, clarifications, and additions)

Interview Questions: (In your opinion...)

1. Purpose of the [name of district removed] Teacher Evaluation System
 - SYSTEM DESIGN AND IMPLEMENTATION (*Organizational*)
 - What is the purpose of the teacher evaluation system?
 - Do you believe teachers share a common understanding of the purpose of the system? Why or why not?
 - To what extent have the processes for designing and implementing the evaluation system at the district-level been transparent?
 - How much input do you feel teachers have had in the design and implementation processes?
 - TEACHER EVALUATION PROCESS (*Individual*)
 - Do you believe that the steps of the teacher evaluation process have been clearly defined?
 - Do you feel school administrators are well prepared to evaluate teachers? Why or why not?
 - What part of the teacher evaluation process do you value most? Why?
2. Measuring Teacher Quality
 - Do you believe the Danielson FFT measures the most important aspects of teacher quality? Are there any domains or components you think are missing?
 - Do you believe there is consistency among evaluators across the district? Why or why not?
 - Is the teacher evaluation process applied fairly to all teachers?
 - Do you believe the value-added model is a good measure of teacher quality? Why or why not?

- If teachers' professional practice scores and value-added scores are not aligned, in which would you place more confidence? Why?
- In general, do you believe that your final classification/label (based on the pilot year data) reflects your level of effectiveness?

3. Impact on Professional Practice

- How has participation in the teacher evaluation process impacted you professionally and personally?
- Has it changed your professional practice? If so, in what ways?
- Has the evaluation process impacted the professional practice of teachers of other teachers at your school thus far? If so, in what ways?
- What impact, if any, do you think teachers' final classifications/labels will have on teacher hiring and retention at your school?
- What impact, if any, do you think teachers' final classifications/labels will have on the perceptions of parents, students, and others in the community?

4. Improving Implementation

- In what ways, if any, can teacher evaluation be improved?
- How can the teacher evaluation system as a whole be improved?
- How can the evaluation process be improved?
- What additional training, if any, would be helpful for you in terms of the Danielson FFT rubric or the overall teacher evaluation process?

5. Additional Comments

- Is there anything you would like to add?

APPENDIX K

SCHOOL ADMINISTRATOR SURVEY PROTOCOL

Part 1: Educator Position

This section includes questions about your current position.

Which of the following best describes your position?

- a. Elementary Principal
- b. Elementary Assistant Principal
- c. High School Principal
- d. High School Assistant Principal
- e. Other, please explain:

Do you work in a Title 1 school?

- a. Yes
- b. No

Part 2: Purpose of Teacher Evaluation

This section includes questions about the purpose of implementing/conducting evaluations of teacher professional practice.

1. In your opinion, what is the primary reason the District evaluates the professional practice of teachers? (*select only one*)
 - a. To help improve the quality of their professional practice to become better teachers
 - b. To hold teachers accountable for their practices/performance
 - c. To make tenure/employment decisions (e.g., improvement plans, termination, reassignment, etc.)
 - d. Mostly to comply with state legislation
 - e. Other, please explain:

2. In your opinion, what should be the primary reason for evaluating the professional practices of teachers? (*select only one*)
 - a. To help improve the quality of their professional practice to become better teachers
 - b. To hold teachers accountable for their practices/performance
 - c. To make tenure/employment decisions (e.g., improvement plans, termination, reassignment, etc.)
 - d. Mostly to comply with state legislation
 - e. Teachers should not be evaluated
 - f. Other, please explain:

Part 3: Content Adequacy of Evaluation Measures

This section includes questions about the content adequacy of measures of teacher quality.

3. To what extent do the twenty-two components of the Danielson Framework for Teaching (FFT) incorporate all/most of the important characteristics of a good/effective teacher?
 - a. The FFT covers all/most of the important characteristics of good/effective teaching.
 - b. The FFT covers some of the important characteristics, but there are a few important attributes/indicators still missing that should be added.
 - c. The FFT covers only a few of the important characteristics of good/effective teaching; there are many attributes/indicators that should be added.

4. What, if any, important attributes/characteristics of good/effective teaching do you feel should be added to the evaluation system? Please explain:

5. When evaluating teachers, which of the following components provides the best indication of what it means to be good/effective?
 - a. Professional Practice: Professional practices ratings on the Danielson Rubrics
 - b. Student Achievement: Measures of student academic achievement (i.e., value-added score, growth score, and/or some type of test score)
 - c. Combination: A combination of both Danielson ratings and student achievement measures
 - d. Neither: Neither Danielson ratings nor student achievement measures

6. A teacher’s overall evaluation score is currently computed as a combination of two primary factors: Danielson FFT rubric ratings (67%) and growth in student achievement (33%). In your opinion, how much weight should be given to each of these components and to any additional components you believe should be represented?

To answer, please enter a percentage between 0 and 100 for each component below. If additional factors should be considered, be sure to include the corresponding percentage weight. Your responses should add up to 100.

Danielson FFT (Ratings of Professional Practice) _____
 Student Achievement (Growth) Measures _____
 Other _____
 Other _____

7. Please explain what you meant by “other” and/or your rationale for assigning each of these weighting factors:

8. Non-Test Information: Should the district consider adding any of the following non-test information to the evaluation criteria? (*Scale: Yes, this should definitely be considered; Possibly; No, this should not be considered*)

- a. Parents: Measures of parent satisfaction with teacher/school
 - b. Students: Measures of attitude, satisfaction, connection with teacher/school
 - c. Peers: Peer-based feedback of teacher/school quality
 - d. Other, please explain:
9. Alternative Achievement Measures: Should the district consider adding additional types of student achievement/learning measures to the evaluation system? (*Scale: Yes, this should definitely be considered; Possibly; No, this should not be considered*)
- a. District benchmark assessments
 - b. Formative measures of student learning
 - c. Performance-based assessments (e.g., projects, portfolios, work samples, etc.)
 - d. End-of-course assessments
 - e. Course grades/Grade point average (GPA)
 - f. Other subject areas (e.g., science, social studies, fine arts, etc.)
 - g. School-wide college-ready indicators (e.g., ACT, SAT, Advanced Placement, etc.)
 - h. School-wide dropout-graduation rates (high school only)
 - i. Other, please explain:
10. I believe that the Overall Effectiveness Classification Labels teachers received for the 2012-2013 pilot year were an accurate representation of their professional performance (i.e., Highly Effective, Effective, Developing, Ineffective).
- a. Strongly Agree
 - b. Agree
 - c. Disagree
 - d. Strongly Disagree

Part 4: Teacher Evaluation System Components

This section includes questions about the components of the [name of district removed] Teacher Evaluation System.

11. During this school year (2013-2014), with what proportion of teachers at your school were the following activities conducted/completed as part of their evaluation? (*Scale: All/Nearly All, Some, Few/None*)
- a. Personal Self-Assessment
 - b. Individual Professional Development Plan
 - c. Beginning of the Year Conference
 - d. Walk-through Observation(s)
 - e. Informal Observation(s)
 - f. Pre-Conference(s)
 - g. Formal Observation(s)

- h. Reflection on Formal Observation(s)
- i. Post Conference(s)
- j. End of Year Conference (*check this if most have already been held, have been scheduled, or will be scheduled prior to the end of the school year*)

12. During this school year (2013-2014), how useful have each of the following evaluation activities been in helping teachers at your school improve their professional practice? (*Scale: Was not conducted as part of the evaluation process, Very useful, Somewhat useful, Not very useful*)
- a. Discussion about their Personal Self-Assessment
 - b. Discussion about their Individual Professional Development Plan
 - c. Discussion/feedback in their Beginning of the Year Conference
 - d. Feedback provided from Walk-through Observation(s)
 - e. Feedback provided from Informal Observation(s)
 - f. Discussion/feedback provided during their Pre-Conference(s)
 - g. Feedback received from Formal Observation(s)
 - h. Discussion/feedback provided via Reflection on Formal Observation(s) during Post Conference(s)
 - i. Discussion/feedback received during their End of Year Conference (*if applicable at this point in the school year*)

Please indicate your level of agreement with the following statements: (*Scale: Strongly Agree, Agree, Disagree, Strongly Disagree*)

13. I feel very comfortable explaining to a non-educator how a teacher's ...
- a. Professional practice (Danielson) score is calculated.
 - a. Value-added (student growth) score is calculated.
 - b. Performance Group Assignment is determined.
 - c. Overall Effectiveness Classification (i.e., Highly Effective, Effective, Developing, Ineffective) is determined.
14. I believe that teachers have control over, and can improve, their future...
- a. Professional practice (Danielson) score
 - b. Value-added (student growth) score
 - c. Overall Effectiveness Classification (i.e., Highly Effective, Effective, Developing, Ineffective)

Part 5: Measuring Teacher Effectiveness

This section includes questions about the overall fairness of the [name of district removed] Teacher Evaluation System.

15. Please indicate your level of agreement with the following statements: (*Scale: Strongly Agree, Agree, Disagree, Strongly Disagree*)

- a. I believe that I am able to evaluate teachers in an objective and unbiased manner.
- b. I believe that I have been well trained in the use of the Danielson rubrics to evaluate teachers.
- c. I believe that I have been able to spend enough time in teacher's classrooms (or professional settings) to adequately evaluate them.
- d. I believe that the evaluation system accurately captures the impact teachers have on improving student motivation, attitudes, and engagement in the learning environment.
- e. I believe that the evaluation system adequately takes into account (adjusts for) the influence of student background characteristics (i.e., demographics, prior achievement, program membership – special education, English language learner, gifted, eligible for free or reduced lunch) when determining teacher's level of professional performance.
- f. I believe that the evaluation system fairly measures the instructional/professional quality of teachers in Group A (using classroom-level value-added [growth] data for their students).
- g. I believe that the evaluation system fairly measures the instructional/professional quality of teachers in Group B (using school-level value-added [growth] data).

16. Improvements: The teacher evaluation system would be significantly improved if... *(Scale: Strongly Agree, Agree, Disagree, Strongly Disagree)*

- a. Teachers were evaluated by more than one observer (not solely their administrator).
- b. Teachers were evaluated by an expert in their instructional content area.
- c. Teachers were evaluated using external evaluators (external to their school).
- d. Teachers were evaluated (in part) by peer-evaluators (other teachers).
- e. Administrators received more training on the Danielson rubrics.
- f. Teachers received more training on the Danielson rubrics.
- g. The Danielson rubric criteria were clarified or better defined.

17. Improvements: In what ways, if any, could the teacher evaluation system or its implementation be improved?

18. Professional Development/Information: What additional professional development, training, or information (if any) related to the teacher evaluation system would be beneficial for you?

Part 6: Evaluation Implementation/Communication

This section includes questions related to communication and the implementation of the [name of district removed] Teacher Evaluation System.

19. Communication: How well has the district informed/communicated with you regarding the development and implementation of the teacher evaluation system?
- Very well
 - Adequately
 - Not very well
20. Resources: How helpful were the following resources in improving your understanding of the purpose, design, and processes of the teacher evaluation system? (*Scale: Very helpful, Somewhat helpful, Not very helpful, Never accessed*)
- Online: Resource links on the district website
 - Online: Videos featured on the district website
 - Online: Comprehensive Evaluation System (CES)/resources
 - Online: [name of district removed] Teacher Evaluation Handbook
 - Professional Development: Led by district administrators
 - Professional Development: Led by the instructional growth teacher or other staff at your school
 - Formal/informal communication: with district administrators
 - Formal/informal communication: with school administrators
 - Formal/informal communication: with a member of the Teacher Evaluation Committee
 - Formal/informal communication: with teachers at your school
21. Classification Report: Please indicate your level of agreement with the following statements about the Teacher Effectiveness Classification Report teachers recently received (*Scale: Strongly Agree, Agree, Disagree, Strongly Disagree, I have not seen this report*)
- Overall, the report included all the important information about the teacher evaluation.
 - The descriptions for each section of the report were easy to understand.
 - The additional resource links helped me better understand the teacher evaluation.
 - I still had questions about how the teacher classification was determined after reading the report.

Part 7: Impact of Teacher Evaluation

This section includes questions regarding the impact (or potential impact) of the [name of district removed] Teacher Evaluation System on school administrators, teachers, and students.

22. Overall, has the evaluation system had a positive or negative impact on your instructional/professional practices?
- Generally positive
 - No real impact

- c. Generally negative

23. In what way(s) has the evaluation system impacted your professional practice?

(select all that apply)

- a. Provided clarity and focus on important aspects of good/effective teaching
- b. Prompted me to be more reflective of my practices
- c. Raised my level of stress/apprehension
- d. Created dialogue, communication, discussion about good/effective teaching practices with teachers at my school
- e. In a negative way, I have narrowed my evaluation of teachers to just what is being evaluated under the Danielson Framework
- f. In a positive way, I have focused my evaluation of teachers to what is being evaluated under the Danielson Framework
- g. Increased my focus on College and Career Ready Standards
- h. Increased my focus on the importance of AIMS test scores
- i. Other, please explain:

24. Overall, has the evaluation system had a positive or negative impact on the professional practices of teachers at your school?

- a. Generally positive
- b. No real impact
- c. Generally negative

25. In what way(s) has the evaluation system impacted the professional practice of teachers at your school? (select all that apply)

- a. Provided clarity and focus on important aspects of good/effective teaching
- b. Increased their use of innovative instructional techniques/activities
- c. Reduced/Limited their use of innovative instructional techniques/activities
- d. Prompted them to be more reflective of their practices
- e. Raised their level of stress/apprehension
- f. Created dialogue, communication, discussion about good/effective teaching practices with me or other school administrators
- g. In a negative way, they have narrowed their professional practices to just what is being evaluated under the Danielson Framework
- h. In a positive way, they have focused their professional practices to what is being evaluated under the Danielson Framework
- i. Increased their focus on College and Career Ready Standards
- j. Limited their instruction to just what is tested on AIMS
- k. Enhanced their focus on individualized student instruction
- l. Improved their communication with parents
- m. Other, please explain:

26. Overall, has the evaluation system had a positive or negative impact on student academic achievement and learning?

- a. Generally positive

- b. No real impact
- c. Generally negative

27. If there is anything else that you would like to add about the impact/consequences of the teacher evaluation system, please do so here:

Part 8: Demographics

This section includes questions related to demographics. Data collected will be used to confirm that respondents are representative of administrators across the district.

28. What is your gender?

- a. Male
- b. Female

29. With which of the following race/ethnicity demographics do you most identify?

- a. American Indian
- b. Asian
- c. Black or African American
- d. Hispanic or Latino
- e. Native Hawaiian or Pacific Islander
- f. Two or more
- g. White
- h. Other

30. Including this year, how many total years have you been an administrator (including experience outside the District)?

- a. 1-3
- b. 4-6
- c. 7-9
- d. 10-12
- e. 13-15
- f. 16 or more

31. Including this year, how many years have you been an administrator in the District?

- a. 1-3
- b. 4-6
- c. 7-9
- d. 10-12
- e. 13-15
- f. 16 or more

Thank you for your participation!

APPENDIX L
TEACHER SURVEY PROTOCOL

Part 1: Educator Position

This section includes questions about your current position.

Which of the following best describes your position?

- a. Classroom Teacher (non-special education)
- b. Classroom Teacher (special education)
- c. Certified Support Position (e.g., Instructional Growth Teacher, Gifted Specialist, Special Education Lead, Data Specialist)
- d. K-8 Counselor
- e. High School Counselor
- f. Related Services (e.g., Speech Therapist, Occupational Therapy/Physical Therapy, Visually Impaired, Hearing Impaired)
- g. Other, please explain:

Do you work in a Title 1 school?

- a. Yes
- b. No

Currently, in which grade level(s) do you primarily teach?

- a. PreK – Grade 2
- b. Grades 3-6
- c. Grades 7-8
- d. All Grades (K-8)
- e. High School
- f. Special Education
- g. Other, please explain:

Are you currently in Group A or Group B?

Group A: K-8 teachers who directly teach state standards in reading, math, and/or science and whose students take the AIMS Reading, Math, and/or Science tests (test scores are directly aligned to what you teach)

Group B: All high school and any K-8 teachers who do not meet the qualifications for Group A

- a. Group A
- b. Group B
- c. Unsure

Part 2: Purpose of Teacher Evaluation

This section includes questions about the purpose of implementing/conducting evaluations of teacher professional practice.

1. In your opinion, what is the primary reason the District evaluates the professional practice of teachers? (*select only one*)
 - a. To help improve the quality of their professional practice to become better teachers
 - b. To hold teachers accountable for their practices/performance
 - c. To make tenure/employment decisions (e.g., improvement plans, termination, reassignment, etc.)
 - d. Mostly to comply with state legislation
 - e. Other, please explain:

2. In your opinion, what should be the primary reason for evaluating the professional practices of teachers? (*select only one*)
 - a. To help improve the quality of their professional practice to become better teachers
 - b. To hold teachers accountable for their practices/performance
 - c. To make tenure/employment decisions (e.g., improvement plans, termination, reassignment, etc.)
 - d. Mostly to comply with state legislation
 - e. Teachers should not be evaluated
 - f. Other, please explain:

Part 3: Content Adequacy of Evaluation Measures

This section includes questions about the content adequacy of measures of teacher quality.

3. To what extent do the twenty-two components of the Danielson Framework for Teaching (FFT) incorporate all/most of the important characteristics of a good/effective teacher?
 - a. The FFT covers all/most of the important characteristics of good/effective teaching.
 - b. The FFT covers some of the important characteristics, but there are a few important attributes/indicators still missing that should be added.
 - c. The FFT covers only a few of the important characteristics of good/effective teaching; there are many attributes/indicators that should be added.

4. What, if any, important attributes/characteristics of good/effective teaching do you feel should be added to the evaluation system? Please explain:

5. When evaluating teachers, which of the following components provides the best indication of what it means to be good/effective?
 - a. Professional Practice: Professional practices ratings on the Danielson Rubrics

- b. Student Achievement: Measures of student academic achievement (i.e., value-added score, growth score, and/or some type of test score)
 - c. Combination: A combination of both Danielson ratings and student achievement measures
 - d. Neither: Neither Danielson ratings nor student achievement measures
6. A teacher’s overall evaluation score is currently computed as a combination of two primary factors: Danielson FFT rubric ratings (67%) and growth in student achievement (33%). In your opinion, how much weight should be given to each of these components and to any additional components you believe should be represented?

To answer, please enter a percentage between 0 and 100 for each component below. If additional factors should be considered, be sure to include the corresponding percentage weight. Your responses should add up to 100.

Danielson FFT (Ratings of Professional Practice) _____
 Student Achievement (Growth) Measures _____
 Other _____
 Other _____

7. Please explain what you meant by “other” and/or your rationale for assigning each of these weighting factors:
8. Non-Test Information: Should the district consider adding any of the following non-test information to the evaluation criteria? (*Scale: Yes, this should definitely be considered; Possibly; No, this should not be considered*)
- a. Parents: Measures of parent satisfaction with teacher/school
 - b. Students: Measures of attitude, satisfaction, connection with teacher/school
 - c. Peers: Peer-based feedback of teacher/school quality
 - d. Other, please explain:
9. Alternative Achievement Measures: Should the district consider adding additional types of student achievement/learning measures to the evaluation system? (*Scale: Yes, this should definitely be considered; Possibly; No, this should not be considered*)
- a. District benchmark assessments
 - b. Formative measures of student learning
 - c. Performance-based assessments (e.g., projects, portfolios, work samples, etc.)
 - d. End-of-course assessments
 - e. Course grades/Grade point average (GPA)
 - f. Other subject areas (e.g., science, social studies, fine arts, etc.)
 - g. School-wide college-ready indicators (e.g., ACT, SAT, Advanced Placement, etc.)

- h. School-wide dropout-graduation rates (high school only)
- i. Other, please explain:

10. I believe that the Overall Effectiveness Classification Label I received for the 2012-2013 pilot year was an accurate representation of my professional performance (i.e., Highly Effective, Effective, Developing, Ineffective).
- a. Strongly Agree
 - b. Agree
 - c. Disagree
 - d. Strongly Disagree
 - e. I was not evaluated in the 2012-2013 school year

Part 4: Teacher Evaluation System Components

This section includes questions about the components of the [name of district removed] Teacher Evaluation System.

11. During this school year (2013-2014), which of these activities were conducted/completed as part of your evaluation? (*select all the apply*)
- a. I conducted a Self-Assessment
 - b. I developed an Individual Professional Development Plan
 - c. I participated in a Beginning of the Year Conference with my administrator
 - d. Administrator(s) conducted Walk-through Observation(s) in my classroom
 - e. Administrator(s) conducted Informal Observation(s) in my classroom
 - f. I participated in a Pre-Conference with my Administrator
 - g. My administrator completed a Formal Observation(s) of my teaching
 - h. I engaged in Reflection on Formal Observation(s)
 - i. I participated in a Post Conference with my Administrator
 - j. I participated in an End of Year Conference with my administrator (*check this if already held, has been scheduled, or you know it will be scheduled prior to the end of the school year*)
12. During this school year (2013-2014), how useful have each of the following evaluation activities been in helping you improve your professional practice? (*Scale: Was not conducted as part of my evaluation, Very useful, Somewhat useful, Not very useful*)
- a. Personal Self-Assessment
 - b. Completing an Individual Professional Development Plan
 - c. Discussion/feedback in a Beginning of the Year Conference
 - d. Feedback received from Walk-through Observation(s)
 - e. Feedback received from Informal Observation(s)
 - f. Discussion/feedback received during Pre-Conference(s)
 - g. Feedback received from Formal Observation(s)

- h. Discussion/feedback received via Reflection on Formal Observation(s) during Post Conference(s)
- i. Discussion/feedback received during an End of Year Conference (*if applicable at this point in the school year*)

Please indicate your level of agreement with the following statements: (*Scale: I was not evaluated during the 2012-2013 school year, Strongly Agree, Agree, Disagree, Strongly Disagree*)

13. I feel very comfortable explaining to a non-educator how my...
- a. Professional practice (Danielson) score is calculated.
 - d. Value-added (student growth) score is calculated.
 - e. Performance Group Assignment is determined.
 - f. Overall Effectiveness Classification (i.e., Highly Effective, Effective, Developing, Ineffective) is determined.
14. I believe that I have control over, and can improve, my future...
- a. Professional practice (Danielson) score
 - b. Value-added (student growth) score
 - c. Overall Effectiveness Classification (i.e., Highly Effective, Effective, Developing, Ineffective)

Part 5: Measuring Teacher Effectiveness

This section includes questions about the overall fairness of the [name of district removed] Teacher Evaluation System.

15. Please indicate your level of agreement with the following statements: (*Scale: Strongly Agree, Agree, Disagree, Strongly Disagree*)
- a. I believe that my Administrator is able to evaluate teachers in an objective and unbiased manner.
 - b. I believe that my Administrator has been well trained in the use of the Danielson rubrics to evaluate teachers.
 - c. I believe that my Administrator has spent enough time in my classroom (or professional setting) to adequately evaluate me.
 - d. I believe that the evaluation system accurately captures the impact teachers have on improving student motivation, attitudes, and engagement in the learning environment.
 - e. I believe that the evaluation system adequately takes into account (adjusts for) the influence of student background characteristics (i.e., demographics, prior achievement, program membership – special education, English language learner, gifted, eligible for free or reduced lunch) when determining my level of professional performance.

- f. I believe that the evaluation system fairly measures the instructional/professional quality of teachers in Group A (using classroom-level value-added [growth] data for their students).
 - g. I believe that the evaluation system fairly measures the instructional/professional quality of teachers in Group B (using school-level value-added [growth] data).
16. Improvements: The teacher evaluation system would be significantly improved if... (*Scale: Strongly Agree, Agree, Disagree, Strongly Disagree*)
- a. Teachers were evaluated by more than one observer (not solely your administrator).
 - b. Teachers were evaluated by an expert in their instructional content area.
 - c. Teachers were evaluated using external evaluators (external to your school).
 - d. Teachers were evaluated (in part) by peer-evaluators (other teachers).
 - e. Administrators received more training on the Danielson rubrics.
 - f. Teachers received more training on the Danielson rubrics.
 - g. The Danielson rubric criteria were clarified or better defined.
17. Improvements: In what ways, if any, could the teacher evaluation system or its implementation be improved?
18. Professional Development/Information: What additional professional development, training, or information (if any) related to the teacher evaluation system would be beneficial for you?

Part 6: Evaluation Implementation/Communication

This section includes questions related to communication and the implementation of the [name of district removed] Teacher Evaluation System.

19. Communication: How well has the district informed/communicated with you regarding the development and implementation of the teacher evaluation system?
- a. Very well
 - b. Adequately
 - c. Not very well
20. Resources: How helpful were the following resources in improving your understanding of the purpose, design, and processes of the teacher evaluation system? (*Scale: Very helpful, Somewhat helpful, Not very helpful, Never accessed*)
- a. Online: Resource links on the district website
 - b. Online: Videos featured on the district website
 - c. Online: Comprehensive Evaluation System (CES)/resources
 - d. Online: [name of district removed] Teacher Evaluation Handbook

- e. Professional Development: Led by district administrators
- f. Professional Development: Led by the principal, instructional growth teacher, or other staff at your school
- g. Formal/informal communication: with district administrators
- h. Formal/informal communication: with administrators at your school
- i. Formal/informal communication: with a [name of district removed] Education Association representative
- j. Formal/informal communication: with a member of the Teacher Evaluation Committee
- k. Formal/informal communication: with other teachers (peers)

21. Classification Report: Please indicate your level of agreement with the following statements about the Teacher Effectiveness Classification Report you recently received (*Scale: Strongly Agree, Agree, Disagree, Strongly Disagree, I did not receive this report*)
- a. Overall, the report included all the important information about my evaluation.
 - b. The descriptions for each section of the report were easy to understand.
 - c. The additional resource links helped me better understand my evaluation.
 - d. I still had questions about my evaluation after reading the report.

Part 7: Impact of Teacher Evaluation

This section includes questions regarding the impact (or potential impact) of the [name of district removed] Teacher Evaluation System on teachers and students.

22. Overall, has the evaluation system had a positive or negative impact on your instructional/professional practices?
- a. Generally positive
 - b. No real impact
 - c. Generally negative
23. In what way(s) has the evaluation system impacted your professional practice? (*select all that apply*)
- a. Provided clarity and focus on important aspects of good/effective teaching
 - b. Increased my use of innovative instructional techniques/activities
 - c. Reduced/Limited my use of innovative instructional techniques/activities
 - d. Prompted me to be more reflective of my practices
 - e. Raised my level of stress/apprehension
 - f. Created dialogue, communication, discussion about good/effective teaching practices with my administrator/peers
 - g. In a negative way, I have narrowed my professional practices to just what is being evaluated under the Danielson Framework
 - h. In a positive way, I have focused my professional practices to what is being evaluated under the Danielson Framework

- i. Increased my focus on College and Career Ready Standards
- j. Limited my instruction to just what is tested on AIMS
- k. Enhanced my focus on individualized student instruction
- l. Improved my communication with parents
- m. Other, please explain:

24. Overall, has the evaluation system had a positive or negative impact on student academic achievement and learning?

- a. Generally positive
- b. No real impact
- c. Generally negative

25. If there is anything else that you would like to add about the impact/consequences of the teacher evaluation system, please do so here:

Part 8: Demographics

This section includes questions related to demographics. Data collected will be used to confirm that respondents are representative of educators across the district.

26. What is your gender?

- a. Male
- b. Female

27. With which of the following race/ethnicity demographics do you most identify?

- a. American Indian
- b. Asian
- c. Black or African American
- d. Hispanic or Latino
- e. Native Hawaiian or Pacific Islander
- f. Two or more
- g. White
- h. Other

28. Including this year, how many total years have you been an educator (including experience outside the District)?

- a. 1-3
- b. 4-6
- c. 7-9
- d. 10-12
- e. 13-15
- f. 16 or more

29. Including this year, how many years have you been an educator in the District?

- a. 1-3

- b. 4-6
- c. 7-9
- d. 10-12
- e. 13-15
- f. 16 or more

Thank you for your participation!

APPENDIX M

SCHOOL ADMINISTRATOR SURVEY PARTICIPATION LETTER

Dear School Administrator,

The [name of district removed] is conducting an ongoing evaluation of the district's Teacher Evaluation System. This survey is being administered to educators throughout the district. The purpose is to obtain feedback to help decision makers improve the future effectiveness of the evaluation system. Aggregate survey results will be reported to and used by district leadership, the District Teacher Evaluation Committee, and other stakeholders. As an administrator in this district, you are invited to participate. The survey should take approximately 20-30 minutes to complete. The survey is divided into eight sections:

Part 1: Educator Position

Part 2: Purpose of Teacher Evaluation

Part 3: Content Adequacy of Evaluation Measures

Part 4: Teacher Evaluation System Components

Part 5: Measuring Educator Effectiveness

Part 6: Evaluation Implementation/Communication

Part 7: Impact of Teacher Evaluation

Part 8: Demographics

Voluntary Participation:

Please note participation in this program evaluation process is completely voluntary. There are no foreseeable risks related to your participation.

Confidentiality:

All of the information that provided is strictly confidential. Neither personally identifiable information nor the name of your school will be asked in the survey. All data/information will be combined with responses from other participants and analyzed/reported in aggregate form.

Please DO NOT mention a person or school name in any comments you provide. The [name of department removed] will redact references to names or schools in all reports.

Only [name of department removed] will have access to the data for collecting, processing, and summarizing the information in a district level evaluation report. In addition, portions of this research may be used in an academic dissertation (Arizona State University Institutional Review Board ID: STUDY00000467), professional reports, conferences/presentations, or related publications.

Questions:

If you have any questions or concerns, please feel free to contact me directly. Thank you for your time and consideration.

Sincerely,

Noelle A. Paufler

APPENDIX N

TEACHER SURVEY PARTICIPATION LETTER

Dear Educator,

The [name of district removed] is conducting an ongoing evaluation of the district's Teacher Evaluation System. This survey is being administered to educators throughout the district. The purpose is to obtain feedback to help decision makers improve the future effectiveness of the evaluation system. Aggregate survey results will be reported to and used by district leadership, the District Teacher Evaluation Committee, and other stakeholders. As an educator in this district, you are invited to participate. The survey should take approximately 20-30 minutes to complete. The survey is divided into eight sections:

Part 1: Educator Position

Part 2: Purpose of Teacher Evaluation

Part 3: Content Adequacy of Evaluation Measures

Part 4: Teacher Evaluation System Components

Part 5: Measuring Educator Effectiveness

Part 6: Evaluation Implementation/Communication

Part 7: Impact of Teacher Evaluation

Part 8: Demographics

Voluntary Participation:

Please note participation in this program evaluation process is completely voluntary. There are no foreseeable risks related to your participation.

Confidentiality:

All of the information provided is strictly confidential. Neither personally identifiable information nor the name of your school will be asked in the survey. All data/information will be combined with responses from other participants and analyzed/reported in aggregate form.

Please DO NOT mention a person or school name in any comments you provide. The [name of department removed] will redact references to names or schools in all reports.

Only [name of department removed] will have access to the data for collecting, processing, and summarizing the information in a district level evaluation report. In addition, portions of this research may be used in an academic dissertation (Arizona State University Institutional Review Board ID: STUDY00000467), professional reports, conferences/presentations, or related publications.

Questions:

If you have any questions or concerns, please feel free to contact me directly. Thank you for your time and consideration.

Sincerely,

Noelle A. Paufler

APPENDIX O

SURVEY PARTICIPATION REMINDER LETTER

Dear _____,

This is a friendly reminder to participate in a research study regarding the new teacher evaluation system in place in the district. Your participation in this survey is very important as findings will directly inform decision making and help improve [name of district removed] Teacher Evaluation System implementation processes.

Please take 10-15 minutes and participate! [Click here to begin:](#)

Please note that your participation is voluntary, and confidentiality will be maintained. If you have any questions concerning this research study, please feel free to contact me directly.

Thank you in advance for your time and help.

Noelle A. Paufler

APPENDIX P
CODE SHEETS

