©2014

Timothy L. Nordin

ALL RIGHTS RESERVED

FEEDBACK FILTER:

EXPLORING FACTORS AFFECTING TEACHERS' USE OF OBSERVATIONAL DATA IN TEACHER EVALUATION

By

TIMOTHY L. NORDIN

A Dissertation submitted to the

Graduate School-New Brunswick

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Education

written under the direction of

William A. Firestone

and approved by

_____

_____

_____

_____

New Brunswick, New Jersey

October, 2014

ABSTRACT OF THE DISSERTATION

Feedback Filter: Exploring Factors Affecting Teachers' Use of Observational Data in Teacher Evaluation

By TIMOTHY L. NORDIN

Dissertation Director:

William A. Firestone

With the advent of a time of intense focus on the use of teacher evaluation in the United States as a mechanism for both formative development of teachers and summative actions affecting their careers, schools, districts, and states have rushed to implement new, enhanced systems of teacher observation and evaluation throughout the country. As many of these new programs are in their infancy, little attention has been paid to how the data generated within these new systems are being used, and what factors affect the usability of such data. While the research literature has focused on data use in the form of data driven decision making in other areas of education (see, e.g., Marsh, Pane & Hamilton, 2006) as well as on improving the reliability and validity of teacher evaluation (see, e.g., Bill and Melinda Gates Foundation, 2013), studying how teachers use evaluation data is a relatively unexplored field.

Using survey and qualitative focus group data from one state's teacher evaluation pilot program, this study explores the ways in which teachers participating in a new evaluation system perceive their data use from observations. Additionally, teacher perceptions around how a number of affective factors –such as time, observer capacity, trust, and data quality, among others –play a role in influencing their ability to use data are explored. These questions help to support the conceptual framework and theory of action presented in the study.

The study finds that teachers see themselves using data in three main ways: to improve practice, to learn about the system, and non-use of data. Additionally, teachers believed that many system functions, as well as data quality issues, were important factors that facilitated or hindered their ability to use the data delivered to them from their observation sessions. The study concludes by developing several hypothetical relationships between these factors and teacher data use, as well as acknowledging study limitations and avenues for further research.

DEDICATION

To Terri,

who never quit, and never let me do so

TABLE OF CONTENTS

ABSTRACT OF DISSERTATION ii

DEDICATION iv

LIST OF FIGURES viii

LIST OF TABLES ix

CHAPTER 1: INTRODUCTION 1

Two Potential Purposes of Teacher Evaluation 3

The Problem of Formative Data Use: The Black Box of Process 5

Research Overview 6

Research Questions 9

CHAPTER 2: LITERATURE REVIEW 11

Conceptual Framework 11

The Changing Face of Teacher Evaluation 15

Influencers of Data-Use in Enhanced Evaluation 22

CHAPTER 3: RESEARCH METHODS 39

The Teacher Evaluation Pilot Program Assessment 39

Research Methodology 46

CHAPTER 4: FINDINGS 60

Perceived Usefulness of Data 61

Three Types of Data Use.....65

Data Use to Change Practice.....66

Data Use to Defend/Increase Evaluation Scores.....69

Data Non-Use.....77

Data Use Summary.....81

CHAPTER 5: FINDINGS 2.....83

System Functionality.....83

Data Quality.....105

CHAPTER 6: DISCUSSION.....120

Summary of Findings.....120

Theoretical Implications.....127

Implications for Policy and Practice.....132

Study Limitations.....137

Avenues for Future Research.....139

Conclusion.....141

APPENDICES.....143

Appendix A: Year 1 Teacher Survey.....143

Appendix B: Year 2 Teacher Survey.....165

Appendix C: Year 1 Interview Guide.....194

Appendix D: Year 2 Focus Group Interview Guide.....195

REFERENCES 199

LIST OF FIGURES

Figure 1: The Skeleton Framework of Teacher Evaluation Date Use…..……6

Figure 2: Conceptual Framework and Theory of Action…..……...16

LIST OF TABLES

Table 1: Perceived usefulness of the observation system ……...…..62

Table 2: Usefulness of feedback provided under the new observation system ...…….63

Table 3: Teacher perceptions about changes in their practice……...……..……67

Table 4: Changing practice to gather higher observation scores …….……73

Table 5: Cross tabulation of teacher responses to alterations……78

Table 6: Of teachers who reported changing neither their instructional practice nor classroom environment, whether they had changed instructional methods to gather higher ratings…..……..……...…78

Table 7: Content knowledge questions in evaluation……..……..85

Table 8: Teacher perceptions of familiarity with observer…….……...……..90

Table 9: Whether the new observation system takes too much time……..……..….93

Table 10: Frequency of feedback from post-observation conferences…….……96

Table 11: Agreement levels with statements regarding trust…….……..…....100

Table 12a: Tenured teachers' response to potential summative consequences……...104

Table 12b: Non-tenured teachers' response to potential summative consequences ……..104

Table 13: Agreement levels with questions pertaining to capabilities of protocol.…….106

Table 14: Agreement levels with questions pertaining to observation accuracy…….112

Table 15: Agreement levels with questions pertaining to observer training…….…….....115

CHAPTER 1: INTRODUCTION

Accelerated Diagnostics, or, The "Rush to Judgment"

Three decades ago, the release of *A Nation at Risk* (National Comission on Excellence in Education, The, 1983) spurred a paradigm shift in public education contributing to the current era of measurement and accountability. The idea that American students were falling behind their international peers was troubling to many, and a number of efforts to create educational systems focused on "tougher standards" for students (Kohn, 2000) and high stakes testing (Amrein & Berliner, 2002) came to the fore. In recent years, the focus on accountability has expanded, as *No Child Left Behind* (Behind, N. C. L., 2002) placed public schools under direct onus to show improvement among students on standardized tests both in general and across student sub-groups.

Even more recently, teachers have found themselves the focus of the accountability movement. Teacher evaluation has become a highly-focused area of educational practice and research over the past decade, with 27 states launching full implementation of new teacher evaluation programs in the 2013-2014 school year; an additional 10 states will reach full implementation by 2014-2015 (Sawchuck, 2013). As educators and policymakers move rapidly to determine teacher proficiencies, or the lack thereof, researchers in teacher evaluation have been equally active, with large scale efforts to understand and improve evaluation of teachers in New York, Chicago, Cincinnati, Washington, DC, Florida, North Carolina, New Jersey, and many others (Bill and Melinda Gates Foundation, 2010; 2012; 2013; Curtis, 2012; Milanowski & Kimball,

2004; Sartain, Stoelinga, & Krone, 2010; Sartain, Stoelinga, & Brown, 2011; VanTassel-Baska, Quek, & Feng, 2007).

While evaluating teacher performance is not new, federal incentives designed to spur states, districts, and schools into creating more rigorous evaluation programs for their teachers have increased attention on this facet of the educational enterprise (Mandinach & Honey, 2008). Across the country, states, districts and schools have moved quickly to implement new evaluative systems, encouraged by demands and funding from NCLB and now *Race to the To*p for "high-quality" teachers (United States Department of Education Office for Planning Evaluation and Policy Development, 2010). This "rush to judgment" (Toch & Rothman, 2008) has created a great deal of new thought around teacher evaluation.

It is unsurprising that teachers would eventually fall under closer scrutiny. Teachers are an important factor in determining student outcomes (Hanushek, 1997; Rowan, Correnti, & Miller, 2006; Wayne & Youngs, 2003), and the focus on educational outcomes driven by the accountability and measurement paradigm has grown beyond simply looking at student measures. As such, teachers are seen by policy entrepreneurs and policymakers as a lens by which to attempt to alleviate some of the perceived failings of the U.S. educational system. As educators, scholars, and policymakers continue to "tinker toward utopia," (Tyack & Cuban, 1995), teacher evaluation has become a significant target of reform efforts.

Though teacher evaluation has been a part of the educational landscape for over sixty years (Toch & Rothman, 2008), in recent times the methodology and rigor of evaluation practices has been called into question (e.g. Bell, Gitomer, McCaffrey,

Hamre, Pianata, & Qi, 2012; Brandt, Mathers, Oliva, Brown-Sims, & Hess, 2007; Braun, 2005). These criticsims have given rise to a new generation of evaluation protocols, focused on increasing the psychometric validity and reliability of teacher evaluation data, as well as attemtping to support the use of data derived from evaluation in decision-making. Typically, these protocols involve altering elements of traditional teacher evaluation, mainly the observation of teacher practice by a superior such as a principal (Oliva, Mathers, & Laine, 2009) with observation protocols designed to alleviate flaws in the traditional system such as observer bias (Antonioni & Park, 2001) and poor reliability of ratings (Bell, Gitomer, McCaffrey, Hamre, Pianata, & Qi, 2012). The reasoning behind such changes can encompass a number of purposes.

Two Potential Purposes of Teacher Evaluation

Teacher evaluation has been carried out for a range of purposes (Glasman & Heck, 1987; Marzano, 2012). One way to characterize the purpose of teacher evaluation is as an opportunity for teachers and administrators to use the data generated by the evaluation process to make decisions about teaching and teachers, whether formative or summative (Flores, 2010). These two branches of data use may be complementary or the two may compete with one another (Flores, 2010; Nevo, 1994).

Formative evaluation seeks primarily to improve teacher practice (Popham, 1988). In a formatively focused system, evaluation data use seeks to engage internal motivations of teachers by delivering constructive feedback during formal or informal interactions with the evaluator. Formative evaluation is deeply connected to teacher needs (Goldstein, 2007; Popham, 1988), as it is designed as an opportunity for teachers to

develop skills and reinforce positive practices. Though formative evaluation –particularly in modern implementations of teacher evaluation systems –may use psychometrically rigorous methods, the reliance on statistically sound measures is not the primary focus. Instead, this type of evaluation requires insight and evidence that can be conveyed to teachers in a meaningful and useful way that point to areas of professional growth. (Goldstein & Noguera, 2006).

Summative purposes generally focus on controlling teacher practice and ensuring that students are not harmed by ineffective teachers. These programs often rely on external motivators for teachers such as salary incentives for effective teachers and financial disincentives, disciplinary action, or outright removal of ineffective teachers. Some states, such as New Jersey, have reformed teacher tenure laws, opening the door to such actions. As New Jersey Governor Chris Christie put it, previous laws in his state had "no rewards for excellence and no consequences for failure." (Perez-Pena, 2011, p. 4). While in a formative system data use by the teacher, or the teacher in collaboration with the evaluator may be common, a purely summative system would see data use primarily on the side of the evaluators. Summative evaluation systems require accurate, reliable and valid measures in order to justify rewards and sanctions delivered to teachers. Teacher needs or concerns are not priorities of such systems. (Goldstein, 2007; Popham, 1988)

Though it is possible that a given evaluation system may be entirely focused on either the formative or summative data use purpose, the districts studied here all combined a blended purpose for their evaluation systems. Moreover, while each district showed variation in its implementation of the new evaluation programs, it is important to note that the formative purpose of data use tended to predominate (Firestone, Blitz,

Gitomer, Gradinarova-Kirova, Shcherbakov, & Nordin, 2013), though teachers recognized that the system also contained important summative elements (Firestone, Nordin, Shcherbakov, Kirova, & Blitz, 2014). This study focuses on formative data-use by teachers.

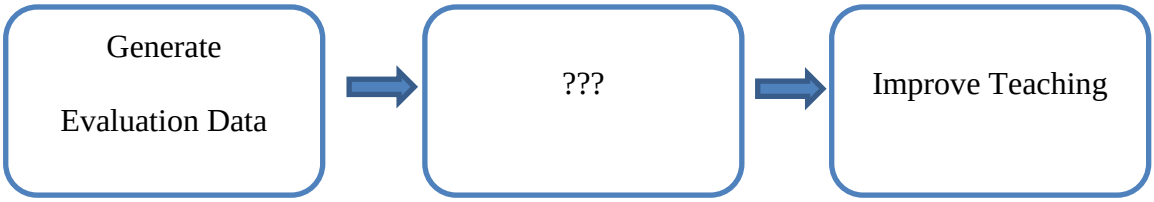
The Problem of Formative Data Use: The Black Box of Process

Under the conception of evaluation as an opportunity for data use, changes in the evaluation systems in public schools are meant to conform to a generalized theory of action which presumes that if teachers are provided constructive data via their evaluation, they will use such data to alter their teaching to become better teachers (Datnow, Park, & Kennedy-Lewis, 2012; Marsh, Pane, & Hamilton, 2006). This ideal, however, may be confounded by a lack of knowledge about several facets of the process of teacher data-use. Though there is a burgeoning literature around the ways in which teachers use student-level data to inform instruction (Hamilton, Halverson, Jackson, Mandinach, Supovitz, & Wayman, 2009; Wohlstetter, Datnow, & Park, 2008), very little is currently known about the actual practice of teacher data use (Coburn & Turner, 2012). Moreover, this work is typically linked to teacher use of student data to drive decisions about curriculum and instruction and has not yet penetrated to studying the use of observational data for formative teacher development.

Additionally, there are signficant gaps in the research literature on understanding how teachers learn (Bakkenes, Vermunt, & Wubbels, 2010); the lack of such knowledge has played a role in the difficulties of many educational innovations to take root (Lieberman & Mace, 2008). These issues make the issue of understanding how teachers

use evaluation data, or even what factors affect teachers use of such data a bit of a black blox. At present, much of the research data use looks something like a simple three step process (*figure 1)*:

Figure 1: The Skeleton Framework of Teacher Evaluation Date Use



There is a need to elaborate the process through which the provision of data contributes to improved teaching to more successfully understand and implement teacher evaluation that produces the desired outcomes. Unfortunately, much of the extant research employs this sort of conceptual framework when investigating data-use (Coburn & Turner, 2012). This study uses data from one state's teacher evaluation pilot program to explore teachers' observational data use and the factors that may affect such data use. Supported by existing literature in the arenas of teacher data use and evaluation, the second step of the skeleton framework will be analyzed more closely, with the aim of generating hypotheses about how the transitions between steps may occur.

Research Overview

This study looks through the lens of those most directly affected by changes in teacher evaluation policies: teachers. Teachers' attitudes, perceptions and needs may color the effects of the evaluation program, though little is understood in terms of about how teachers view these new systems. Past research on educational change suggests that those directly involved in change of any sort play a significant role in shaping the

programs in which they participate (Fullan, 2007; Lipsky, 2010). On a basic level, understanding the feelings and perceptions of teachers and administrators can help facilitate the basic change process (Fullan, 2007). Acting as "street level bureaucrats," (Lipsky, 2010), teachers work to mold new processes to fit with their understandings of the change process, resisting misunderstood or poorly developed steps in the system.

Indeed, successful implementation cannot ignore the individuals who do the implementation (Fullan, 2007), especially if the outcomes are hoped to be formative of teachers. As noted by Hargreaves (2005), "…understanding how teachers experience and respond to educational change is essential if reform and improvement efforts are to be more successful and sustainable" (p. 981). The research literature on educational change delineates many factors that contribute to understanding the change process, from identifying the various stages of adjusting to change (Hall & Hord, 1987) to the emotional and experiential factors affecting change initiatives (Leithwood, 1992). Strong methodology, reliable and accurate measures, and the best –or worst –intentions may be easily derailed by failure to anticipate the ways in which the program may be altered by those actually carrying it out.

Thus, teachers are a vital part of the educational change process, and their interaction with changing evaluation policy is important to generating ideas about evaluation data-use. Therefore, an understanding of how teachers perceive and process not only the data provided to them during evaluation but how various factors are seen as affecting their use of that data may be useful to suggest hypotheses for how the process of data-use and teacher learning occur.

This study approaches the teacher data use in evaluation by using the existing literature on data use and evaluation to develop a conceptual framework and theory of action that describes how teacher data use may be affected by factors that influence strong evaluation techniques. Within the literature review, a fleshed out conceptual framework is presented that effectively reverse-engineers the three step skeleton framework. Beginning from the theory of action implicit in data-use efforts in education and teacher evaluation programs (Datnow, Park, & Kennedy-Lewis, 2012), the existing literature is extended from teacher use of student data to focus on evaluation data use. On the other end of the skeleton framework, the literature review grounds the generation of evaluation data step in two separate literatures supporting several factors that affect teacher data use. One body of work deals with factors of system functionality, drawn from Marsh (2012). In addition to systemic factors, research on effective evaluation techniques that have recently been developed to combat psychometric data quality issues and are derived from a wide swath of educational researchers (see, e.g., Bell, Gitomer, McCaffrey, Hamre, Pianata, & Qi, 2012; Marzano, 2012; Toch & Rothman, 2008).

Drawing from the extant body of literature, this study uses a secondary analysis of data gathered during the assessment of a teacher evaluation pilot program to develop tentative hypotheses about how factors known to affect evaluation may affect the process of teachers' use of evaluation data. Teacher surveys from teachers in districts participating in the pilot program as well as qualitative focus group interviews with teachers are used to explore teachers' perceptions about how they are –or are not –using evaluation data and how systemic and data-quality issues affect teachers' perceived ability to use this data.

Teacher perceptions around these issues led, in some cases, to either a diminishment of the data-use process and, in others, an amplification of the process. The views of teachers suggest that these factors contribute to a potential "feedback filter," by which evaluation data use is moderated –either positively or negatively –by other factors involved in implementing a new evaluation program. Within the methodological discussion, the tentative nature of hypotheses developed by the analysis is addressed; findings are used to note potential areas for further, more focused study to validate the hypothetical model.

Research Questions

The research questions guiding the analysis of the data derive from both the existing literature as well as the lack of research exploring how teachers' data use is filtered through mediating factors. The main research questions are as follows:

1) How do teachers perceive the usefulness of the observational data received in the new evaluation systems?
    a) In what ways, if any, do teachers report using evaluation data?
    b) What data use actions are most commonly reported as important to teachers?
2) How do teachers perceive the effects of known factors influencing evaluation quality and data use?

Following the analysis of the research data, the discussion chapter focuses on developing hypotheses regarding how the factors proposed as affecting data-use may indeed be related to the theoretical outcomes of data-use. Given the tentative nature of these hypotheses, as this study constitutes a secondary, qualitatively focused analysis of

the data collected for assessment of the one state's pilot teacher evaluation program, several suggestions for further research designed to test the tentative hypotheses are also proposed.

Though exploratory in nature, efforts to begin to understand the process of teacher data use in evaluation may prove useful in adding to the knowledge base around evaluation data-use. Understanding the way that teachers filter and use the feedback data provided to them during evaluation may allow parties involved to tailor evaluation plans for more effective outcomes. Given the state of the research body in this area, theory generation is a necessary step to advancing the field and encouraging further research.

In summary, the time is ripe to delve into understanding how teachers use observational data. With many schools, districts, and states across the nation adopting new, more rigorous teacher evaluation programs, the opportunities for teachers to use observational data are now more numerous, in many cases, than ever before. Add to this the increasing pressure on schools and districts to show improvement in student outcomes generated by federal measures and funding opportunities and the need for an understanding of how teachers use data, and how to aid in the use of such data to achieve desired outcomes has never been greater.

CHAPTER 2: LITERATURE REVIEW

Restructure: The Changing Face of Teacher Evaluation and Framing Evaluation Data Use

The body of literature framing this research comes from several areas of inquiry. These include the history of and current efforts in teacher evaluation, factors that affect educational change and data use, and the anticipated outcomes of data-use. Though there is a burgeoning literature for each of these research efforts, there is also a notable lack of focus on these issues in evaluation. This creates an opportunity to begin to enlarge the body of knowledge about the process of evaluation and data-use. This chapter opens with an outlining of the conceptual framework guiding the research, which simultaneously serves as the hypothetical model by which the process of teacher data-use in evaluation may proceed. The remainder of the review of the literature fleshes out the framework by exploring the various elements of the framework.

Conceptual Framework

The use of data to drive decisions about future practice has recently become a focal point of improvement efforts across many disparate fields, including business (i.e. Brynjolfsson, Hitt, & Kim, 2011), medicine (i.e. Sox, Higgins, & Owens, 2013), and education. As defined by Marsh, Pane and Hamilton (2006), data-driven decision making (DDDM) in education "refers to teachers, principals, and administrators systematically collecting and analyzing various types of data… to guide a range of decisions to help improve the success of students and schools" (p. 1). This has been elaborated in many

different ways. Some examples include the work of Wohlstetter, Datnow and Park (2008), who used the *principal-agent theory* (see Eisenhardt, 1989) to explore the effects of autonomy and capacity of educators to apply DDDM. Hamilton and colleauges (2009) delineated much of what is known about the practice of DDDM for teachers. This recent upsurge in data-focused efforts to improve education provides a significant research base upon which to situate the current study.

This dissertation uses the existing literature on DDDM, teacher data-use, and factors affecting evaluation data-use to develop a theoretical model (see figure 1) by which data-use in teacher evaluation may function. Though the research presented in this work will focus on the perceived effects of implementation and practice factors in evaluation, the theoretical model is grounded in the implied theory of action that use of data by agents within the school will produce positive results in both school and student outcomes (Marsh, Pane, & Hamilton, 2006). This concept is supported by Datnow, Park, and Kennedy-Lewis (2012), who note, "The theory of action of [data-driven decision making] policies is that educators need to know how to analyze, interpret, and use data so that they can make informed decisions about how to improve student achievement on state and national assessments" (p. 247). Thus there is a research-based foundation to support the idea that the purpose of data use in education is to provide positive outcomes in the form of student achievement and school improvement.

Drawing on this theory of action, this study extends the idea of positive outcomes to teacher improvement. This is an extension of the literature because teachers can play a significant role in student success (Hanushek, 1997; Hanushek & Rivkin, 2010; Rivkin, Hanushek, & Kain, 2001; Rowan, Correnti, & Miller, 2006; Wayne & Youngs, 2003).

Since teachers can be influential in students' development, ideally, teachers who more effectively use pedagogically sound techniques will see better outcomes for their students. Though much of the research literature on teacher data use focuses on the use of student data to inform instructional decisions, the theory of action is simply and subtly turned to focus on teachers' use of observational data to alter their instruction. In both cases, teachers are the agents of change, using –or failing to use –opportunities to interact with provided data. As the literature suggests, data-use is presumed to lead to positive outcomes. On the other hand, some scholars have noted that there may be unintended consequences that come out of the use of accountability systems to encourage data use (Jennings, 2012; Koretz, 2008). Jennings (2012) refers to these as distortive consequences. These may include changing how lessons are taught compared to a non-observed lesson (Nolan Jr. & Hoover, 2004), and modifying behaviors to attempt to reach higher scores on the evaluation rubric (Firestone, Nordin, Shcherbakov, Kirova, & Blitz, 2014).

There are, however, significant holes in the research literature. There is "shockingly little research on what happens when individuals interact with data in their workplace," (Coburn & Turner, 2012, p. 99). Though there is a clear theory of action to describe the process of data-use providing benefits to students and schools, there remain few studies that address how the process works. As Coburn and Turner point out, those studies that aid in linking data-use programs to outcomes (e.g. Schmoker & Wilson, 1995; Symonds, 2004) do not address the process that teachers undergo to achieve those results. This paucity of research regarding how teachers use data is exacerbated in the arena of teacher evaluation, as only recently have public schools, districts and states

rushed to produce data-driven evaluation systems for teacher improvement (Toch & Rothman, 2008) given the impetus provided by federal measures designed to encourage such actions (Mandinach & Honey, 2008; United States Department of Education Office for Planning Evaluation and Policy Development, 2010).

Therefore, there is a need to expand upon the framework developed by researchers. As yet, there is little research that endeavors to determine how factors of implementation affect data-use in a teacher-evaluation context. Marsh (2012) outlined anumber of factors that affect teacher data use that affect teachers' ability to use student data. These include capacity of the intervener; properties of the data; leadership; time; interpersonal trust and relationships; and data user beliefs, skills, and knowledge, including the organizational context for evaluation. These factors form one portion of the theoretical model from the front end, in this work called "Factors Affecting Teacher Data Use" (*Figure 2)*. Though Marsh's work focused on teacher use of student data, this study adapts the framing of data use used by Marsh and developed by Mandinach, Honey, Light, and Brunner (2008) to include observational data. Here, these factors are called "System Functionality" factors.

In addition to system functionality factors that may affect teachers' ability to use data from evaluation, a number of potential barriers and facilitators to data use are proposed in the literature. These factors fall generally under the umbrella of "Properties of data" (Marsh, 2012, p. 30) and are more specific to the current push toward more psychometrically focused teacher evaluation. The data side of the theoretical model includes factors such as the observation protocols used (e.g. Danielson, 1996; 2007; 2011; Marzano, 2007; Stronge, 2012; Williams, 2009), as well as the reliability and

validity of the evaluation data (Bell, Gitomer, McCaffrey, Hamre, Pianata, & Qi, 2012; Braun, 2005; Toch & Rothman, 2008; Wayne & Youngs, 2003).

A third branch of affective factors concerns characteristics of the teachers themselves. Personal understandings of teaching (Fullan, 2007), comfort with the evaluation instruments, including training in the instruments (Cruickshank & Haefele, 2001; Haefele, 1993; Oliva, Mathers, & Laine, 2009), and understanding of the evaluation criteria by which they are evaluated (Firestone, Blitz, Gitomer, Gradinarova-Kirova, Shcherbakov, & Nordin, 2013).

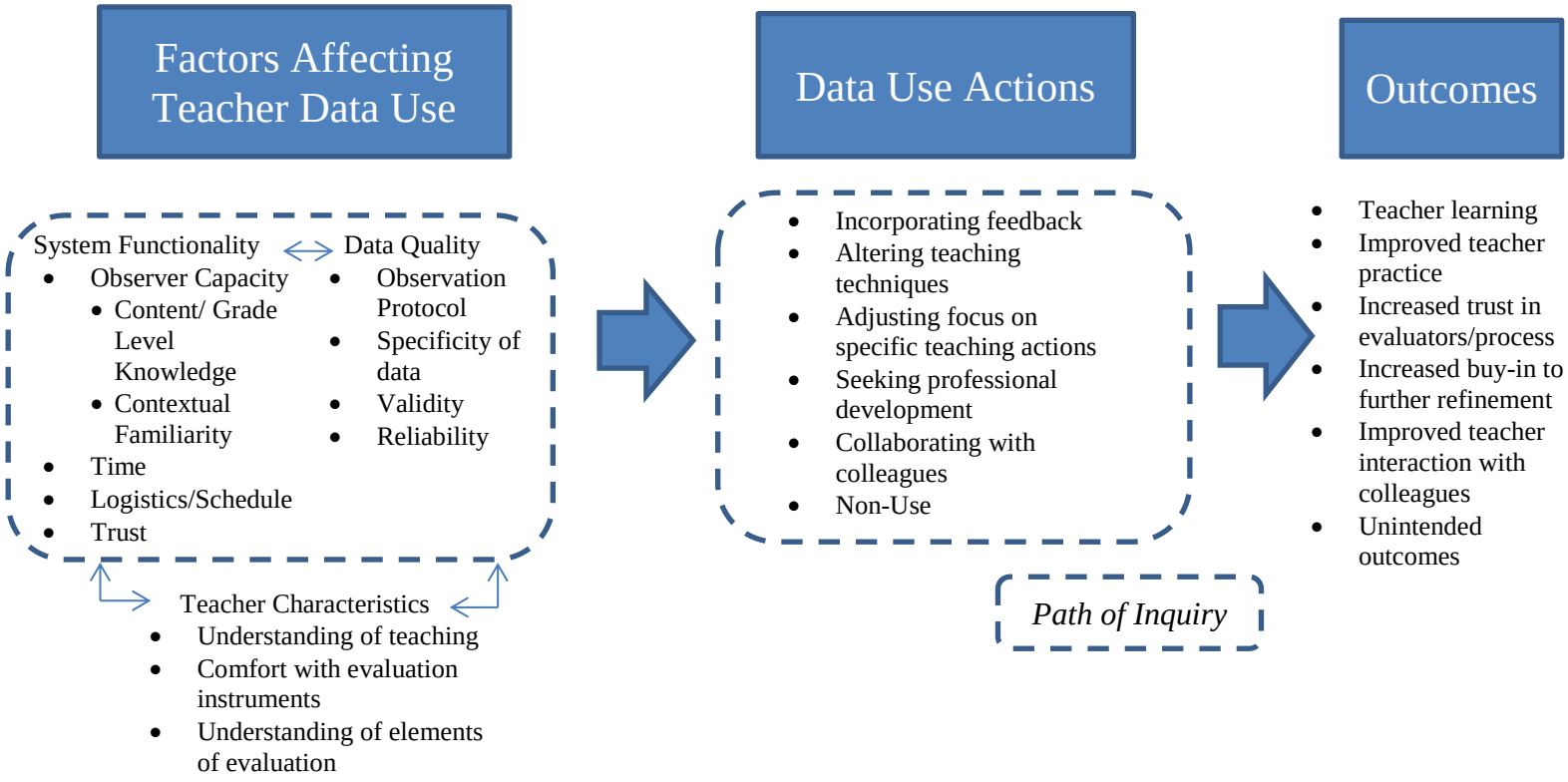
The Changing Face of Teacher Evaluation

These three branches may be important in understanding how data use is affected for teachers in enhanced evaluation programs. But before diving more specifically into these branches of the framework, it is important to outline the current paradigm of teacher evaluation.

Teacher evaluation has been a part of the educational system for well over sixty years (Toch & Rothman, 2008). Recent acknowledgement of the problems of traditional observation has led to new approaches and new studies in an attempt to create a more useful and more psychometrically valid approach.

Figure 2: Conceptual Framework and Theory of Action

Problem: One way to conceptualize the purpose of teacher evaluation is as a formative opportunity for teachers to use evaluation data to teach more effectively. Several factors may contribute to realizing this purpose, though existing research has not yet explored how these factors affect teacher data-use. Without understanding teachers' perceptions of the importance of these factors in their use of the data, one cannot hypothesize how those factors may ultimately affect outcomes for teachers' development.



Problems of Traditional Teacher Evaluation

As Milanowski (2004) stated, "It may seem unusual to think of teacher evaluation systems as a source of information on teacher instructional behavior that affects student learning. As a measurement process, the reputation of teacher evaluation is not particularly good," (p. 34). Marzano (2012) criticizes previous evaluation plans on the grounds of failure to accurately measure teacher quality and failure to develop a "highly skilled teacher workforce," (p. 14).

It is apparent that observation of teachers needs to evolve. As Toch and Rothman (2008) explain, "Because teacher evaluations are at the center of the educational enterprise –the quality of teaching in the nation's classrooms –they are a potentially powerful lever of teacher and school improvement" (p. 1). Toch and Rothman also note that teachers and teacher quality have been typically defined by credentials such as advanced degrees and subject-area certification rather than more statistically valid measures.

Observations by principals have traditionally been the main source of teacher evaluation (Oliva, Mathers, & Laine, 2009; Danielson, 2011). Though research shows principal evaluations to be positively correlated to teachers' value-added measures of student performance (Harris & Sass, 2009) and that principals have an impact on student learning (Hallinger, 2003; Heck & Hallinger, 2009; Leithwood & Jantzi, 1999; 2000; 2005), principal evaluations have a number of weaknesses as evaluative tools.

Observer Bias. Foremost among these challenges is observer bias. In general, relationships with the observer have been shown to cause bias in the ratings (Antonioni & Park, 2001; Robbins & DeNisi, 1994); studies show that principals are often biased in

their evaluation of teachers (e. g. Cash, Hamre, Pianata, & Myers, 2012; Danielson, 2007; Graham, Milanowski, & Miller, 2012). Weisberg and colleagues (2009) and Gitomer and Bell (2013) both found that 99% of teachers are routinely evaluated at the highest levels. In a wry statement, Secretary of Education Arne Duncan said, "Today in our country, 99 percent of teachers are above average," (Gabriel, 2010, p. 3).

As Weisberg and colleagues (2009) point out, there are several major problems with the blatant observer bias displayed in this trend. Since all teachers are considered equally successful, no particular attention is paid to the development of neophyte teachers. Veteran teachers are not tapped for their knowledge and experience. Professional development for all teachers is hindered by the generality of the ratings, as particular strengths and weaknesses are not adequately identified by such ratings. Finally, those teachers just above the bottom 1% receive neither the support needed to remediate their practice nor the motivation to improve. The Widget Effect therefore defeats both summative and formative purposes of education.

Methodological Concerns. In addition to observer bias, observations are methodologically challenging in a psychometric sense. Scholars have criticized the validity of traditional observation instruments (Bell, Gitomer, McCaffrey, Hamre, Pianata, & Qi, 2012), the limited frequency of typical evaluations (Brandt, Mathers, Oliva, Brown-Sims, & Hess, 2007) and the difficulty in attribution of teacher behaviors to student outcomes (Braun, 2005; Darling-Hammond, Amrein-Beardsley, Haertel, & Rothstein, 2012; Darling-Hammond, Wise, & Pease, 1983; Kennedy, 2010; Milanowski, 2004). There is little evidence that these types of evaluations offer either the statistical validity or measurement accuracy needed to make summative decisions about teachers –

awarding tenure or merit-pay, for example –or the formative depth needed to address teacher professional development and improving practice. While both these uses may ultimately benefit student outcomes, it is clear that traditional evaluative methods are insufficient for either purpose.

Despite criticisms, many modern evaluation systems still include teacher observation as a component of a multiple-methodology evaluative program (Bill and Melinda Gates Foundation, 2010, 2012, 2013; New Jersey Educator Effectiveness Task Force, 2011). Though many scholars and policymakers now advocate for the use of external evaluators –that is, evaluators without direct ties to the evaluatee –for rating teachers as well as creating systems to ensure inter-rater reliablity (Bill and Melinda Gates Foundation, 2012; La Paro, Pianata, & Stuhlman, 2004; Little, Goe, & Bell, 2009; Milanowski, 2004), principals and other administrators continue to be common evaluators. This is likely due to many factors, such as the cost of using outside raters, administrators' presumed understanding of the specific educational context of their schools, and/or issues of district capacity.

Designing a New Evaluation Paradigm: External Evaluation and Inter-rater Reliability

The drive for reliable, accurate measurements of teacher performance currently constitutes the move toward a psychometric mindset in evaluation protocols. Increased accuracy and reliability has implications for both formative and summative purposes of evaluation. This is particularly true of the summative purpose; summative actions are defensible only in a paradigm of valid, reliable measures. Without valid measures, rewarding or punishing teachers based on their evaluation performance may be biased,

unjust, and will almost certainly be met with severe resistance (Baker, Oluwole, & Green, 2013; Baker, et al., 2010).

For formative purposes, reliability and accuracy are also desirable traits. By improving the reliability of evaluation measures –teacher observations in particular – teachers may feel more comfortable in receiving feedback from an observer. Modern evaluation protocols, designed with primarily formative purposes in mind (e.g. Danielson C. , 2007; The Danielson Group, 2011; Marzano, 2007; Stronge, 2012), stress the necessity of valid observational scores and reliable measures to enhance teacher performance. These systems aim to incorporate measurement more purposefully into observation.

Addressing the many challenges affecting teacher observation validity has become a primary concern. Many scholars have long suggested that training is essential to accurate observation (Cruickshank & Haefele, 2001; Haefele, 1993; Mathers, Oliva, & Laine, 2008; Oliva, Mathers, & Laine, 2009). Moreover, thorough training of observers has been a required component of both some of the most popular modern observation protocols (Danielson C. , 2007; 2011; 2011) and as a requirement in influential studies of teacher evaluation (Bill and Melinda Gates Foundation, 2013). Training, of course, is a complex undertaking, yet seems to be vital in increasing validity in observation (Cash, Hamre, Pianata, & Myers, 2012). These include the method of training, as Cash and colleagues (2012) found that training turnkey trainers can be successful. Additionally, they found that the beliefs and mindsets of those being trained has a significant effect on outcomes.

Another effort to limit observer bias comes in the form of external observers. Graham, Milanowski and Miller (2012) have pointed out that due to the various biases that affect internal observers, "observers who are less familiar with their ratees will be more likely to rate accurately" (p. 17). Little, Goe, and Bell (2009) argue that external observers remove perceived and real bias in the evaluation process, creating more scientifically valid measurement. The Measures of Effective Teaching (MET) study incorporated entirely external observation by using video-recorded classes and having trained observers rate teachers' effectiveness after the lesson was complete (Bill and Melinda Gates Foundation, 2010, 2012, 2013). The use of external raters has been a part of a number of other observation projects in New Jersey (New Jersey Educator Effectiveness Task Force, 2011), Cincinatti (Milanowski & Kimball, 2004), and in multi-state efforts (La Paro, Pianata, & Stuhlman, 2004), among others.

There are variations in the use of external evaluators. As external observers may be both costly and may miss contextually important factors specific to the classrooms they are observing, some districts have made effort to utilize both internal and external evaluation. In Chicago, principals trained with external raters to observe and evaluate teachers, with positive results concerning principals' ability to achieve high inter-rater agreement with outside evaluators (Sartain, Stoelinga, & Krone, 2010). This evidence suggests that the problems with principal bias seemingly inherent in the observaton process can be mediated with training and ongoing calibration. In Florida, Hillsborough County schools combined principals and peer-raters to observe teachers' practice (Curtis, 2012).

The evidence that principals' biases can be controlled with training and partnership with external observation and that increasing the number of observations increases reliability highlights a growing trend in educational leadership. As principals spend more time observing teachers their roles will naturally grow from building management toward instructional leadership. The instructional leadership paradigm argues that the school's primary purposes are "teaching and learning, and organizing for teaching and learning," (Prestine & Nelson, 2005, p. 46), and as principals and other school leaders spend more time in classrooms observing practice and using observational data to guide decision making, instructional leadership in schools will grow (Knapp, Copland, & Talbert, 2003).

Influencers of Data-Use in Enhanced Evaluation

All this research points to the obvious fact that teacher evaluation needs improvement if it is intended to be used as a serious tool for formative development of teachers, or even for summative actions. As these new systems form and are studied (Bill and Melinda Gates Foundation, 2010; 2012; 2013; Curtis, 2012; Milanowski & Kimball, 2004; Sartain, Stoelinga, & Krone, 2010; Sartain, Stoelinga, & Brown, 2011; VanTassel-Baska, Quek, & Feng, 2007), there is a need for a model of how evaluation systems are meant to affect education. This study uses the existing research literature on data use to frame a hypotheical model for teachers' use of evaluation data. For the purposes of this study, factors that may affect data use are dividided into three main catgories: System Functionality, Data Quality, and Teacher Characteristics. Though this study does not collect data on all the outlined branches of these three catagories, it is important to note

their existence in shaping the potential use of the evaluation data for teachers and overall outcomes.

System Functionality

Factors of system functionality refer mainly to issues that affect the implementation and overall function of the system. The effects of these factors will play out in different ways given the specific context of the site and the specifics of the system, but can be mediated with concerted effort by implementers. These factors flow from the data-use work of Marsh (2012).

Observer Capacity. According to Marsh (2012), intervener capacity "can greatly influence the outcome of [data-use] efforts" (p. 29). In this case, the intervener is defined as the observer of the teacher; in many cases this will be a principal or other administrator. As suggested by many scholars, this may require that the observers possess a robust knowledge of teaching and learning (Knapp, Copland, & Talbert, 2003; Park & Datnow, 2009; Prestine & Nelson, 2005). One way that this has been typically accomplished is via training for the new evaluation system. Training has been found to be a complex yet vital undertaking (Cash, Hamre, Pianata, & Myers, 2012). Many scholars have long suggested that training is essential to accurate observation (Cruickshank & Haefele, 2001; Haefele, 1993; Oliva, Mathers, & Laine, 2009). Moreover, thorough training of observers has been a required component of both some of the most popular modern observation protocols (Danielson C. , 2007; 2011; 2011) and as a requirement in influential studies of teacher evaluation (Bill and Melinda Gates Foundation, 2013).

When discussing the capacity of the observer, two factors are often overlooked in current evaluation reform efforts: contextual and content-specific knowledge. With many

current observation protocols given to generating universal models for evaluating teaching (see, e.g., Danielson C. , 2007; Marzano, 2007; Stronge, 2012; Williams, 2009) the specific context of individual districts, school and classrooms is somewhat ignored. Yet, that context is vital to understanding the teacher's practice. As Stronge and Tucker (1999) put it, "… it becomes crucial that [multiple constituencies'] perspectives are understood and addressed in the evaluation system that is ultimately created." (p. 341). Multiple systems are at play in the classroom, originating from the outside and not necessarily under the control of the teacher (Caro, 2009; Davis-Kean & Sexton, 2009; Rubin, Bukowski, & Parker, 2006). This suggests that contextual understanding of the situation may be vital in properly evaluating a teacher's practice and performance. As Gitomer and Bell (2013) state,

> How much of what we see is under the sole control of the teacher and how much might be attributable to contextual factors that influence what the teacher does and how students learn? For example, while one can judge the quality of the content being taught, that content is frequently influenced by district-imposed curricula and texts. Social interactions that occur among students are certainly a function of the teacher's establishment of a classroom climate, but students also bring a set of interpersonal dynamics into the classroom. (p. 6)

There is evidence, as well, that teachers will see this as an important issue in evaluation. Teachers have been shown to feel vulnerable to facets of their practice that they cannot control (Conley & Glasman, 2008; Yuan, et al., 2013) and thus may have a heightened sensitivity to an observer that they perceive to have little knowledge of the contextual issues affecting their classrooms. Ultimately, this may affect teachers' data-use, as their

actions within the system may change the outcomes (Fullan, 2007; Ginsburg, Kamat, Raghu, & Weaver, 1995).

A second issue of observer capacity is the content knowledge of the observer. Evaluation has a "blind spot with respect to content that characterizes… most of our state-level programs of teacher evaluation and teacher certification," (Shulman, 1986, p. 42). Stated nearly thirty years ago, Shulman's comments hold true today. Though a number of studies and scholars have advanced the use of external observers that may be more content specific than administrators (e.g. Bell, Gitomer, McCaffrey, Hamre, Pianata, & Qi, 2012; Bill and Melinda Gates Foundation, 2013) and principals' responsibilities are shifting toward instructional leadership (Donaldson & Donaldson, Jr., 2012), observation systems and studies built around them continue to rely on definitions of effective teaching that are generalized across grade levels and subject areas.. This is not to argue that the definitions of successful teaching found in these protocols are wrongheaded, simply that they forsake some of the nuance of teaching in a rush to be useful to a larger audience (Ball, Thames, & Phelps, 2008; Shulman, 1986).

At the very least, instructional leadership advocates suggest that leaders should have "leadership content knowledge" in at least one subject (Stein & Nelson, 2003). Leadership content knowledge is defined by Stein and Nelson (2003) as "the kind of knowledge that will equip administrators to be strong instructional leaders," (p. 424). This knowledge need not be as deep as that of the teacher in but must be substantial nonetheless.

Yet content is not the sole difference in the way subject areas are taught. Pedagogical content knowledge (PCK), a term first coined by Shulman (1986) describes

the body of knowledge of how to teach, and how students learn, in a specific subject area. Not only is the content of the physics classroom different from that of the social studies classroom, but *teaching* physics is different than teaching social studies. PCK posits that instruction can only be effective if it is attuned to how students learn (Van Driel & Berry, 2010).

Since teaching across the various grades and subjects differs in both the content itself and the way that content is presented, it seems natural that one must consider the impacts of observer content knowledge on how teachers may use evaluation data. Some scholars, such as Goldstein and Noguera (2006), have advocated for peer assisted reviews, carried out by teachers. This may solve the issue of content knowledge gaps in observers, but is not targeted to psychometrically stable observational data (Goldstein, 2007).

Time. In any given change to an organization, time is an important barrier to address. Hall and Hord (2006) cite time as a common issue affecting systemic implementation of educational change. In their study, Hall and Hord noted that teachers were concerned about "lack of time needed to fully understand the new approach, and lack of time to become involved in the change process and for feedback during the implementation process." (p. 55). Additionally, teachers themselves constantly feel time pressure in their daily work (Fullan, 2007). Huberman (1983) noted that teachers are under a constant press for time, which limits their opportunities for sustained reflection. Having limited time to reflect on the change process can have negative effects on teacher learning, one of the presumed outcomes of the hypothetical model.

Tynjälä (2008) presented seven different ways people learn at work based on a review of research literature: by doing the job itself, through cooperation and interaction among colleagues, through working with clients, by taking on new and challenging tasks, self-reflection and evaluation of one's experiences, formal education, and through experiences outside of work. This framework was used by Bakkenes, Vermunt, and Wubbels (2010) to define teacher learning as, "an active process in which teachers engage in activities that lead to a change in knowledge and beliefs (cognition) and/or teaching practices (behavior)" (p. 536). However, without time to accomplish this, teachers may be unable to successfully adopt new teaching methods or adequately use evaluation data for their formative development. Time is clearly a factor that must be considered in studying teacher perceptions of data use.

Logistics/Schedule. Tied to the issue of time in affecting teacher evaluation data use is the logistics and scheduling of evaluation practices. Since several advocates for change in evaluation recommend the use of multiple measurement points and regular feedback for teachers (Peterson, 2000; Weisberg, Sexton, Mulhern, & Keeling, 2009), scheduling evaluation points becomes a critical factor in creating a system that can be useful to teachers (Danielson & McGreal, 2000). Observing and evaluating teachers over several instances in a school year requires a great deal of time from both observers and teachers. For example, the MET study worked with teachers across 6 participant districts to observe over 20,000 lessons via video recording (Bill and Melinda Gates Foundation, 2010). These lessons were then evaluated by observers of the video recording. Some of the evaluation was done in smaller-than-class-length "lesson segments" which ranged from 7.5 minutes to 15 minutes in length (Bill and Melinda Gates Foundation, 2012).

Even at this short end, teaching and evaluating a 7.5 minutes lesson requires a minimum of 15 minutes of labor. Given that many evaluation protocols suggest or require longer observation times (e.g. Danielson C. , 2011; New Jersey Department of Education, 2013), the number of person-hours increases. This is only the act of observing, and leaves out time required for pre-observation conferencing between teachers and observers as well as post-observation feedback –required by many current observation protocols (see, e.g., Danielson C., 2011; Marzano, 2007; Stronge, 2012; Williams, 2009), the vehicle that provides the main opportunity for teachers to use observational data.

Trust. Within education, the literature has shown that trust is a factor in school improvement on a number of fronts (see Bryk & Schneider, 2003; Hoy & Sweetland, 2001; Louis, 2007; Tschannen-Moran, 2004; Wahlstrom & Louis, 2008). Trust between teachers and observers is essential to the observational process if the purpose of evaluation is formative in nature. Formative evaluation focuses on the development of teachers, and trust between administrators/observers and teachers may enhance the effectiveness of this process. Administrators must demonstrate competence in order to help fulfill this goal (Tschannen-Moran, 2004). Louis (2007) showed that schools that engender a high-trust environment were more likely to have successful reforms that resulted in positive student outcomes. This is obviously essential in a massive, holistic program such as a multiple-measures teacher evaluation program. If nothing else, trust may play a factor in how teachers perceive the program and must be considered in the hypothetical model.

Data Quality

Beyond factors of the evaluation system that may impact teachers' opportunities for data use, the quality of evaluation data has recently been a focus of significant debate and research. A number of recent studies have explored the data quality in order to improve its usability in teacher evaluation (Bill and Melinda Gates Foundation, 2013) or to note the need for higher quality data (Bell, Gitomer, McCaffrey, Hamre, Pianata, & Qi, 2012; Braun, 2005).

Observation Protocol and specificity of data. Two decades ago, Kulm and Stuessy (1991) noted that "changes in curriculum goals (also) require concurrent changes in approaches used by teachers in improving learning" (p. 73). This continues to be true in the current evaluation environment, as observation protocols used to measure teacher success continue to evolve. Two main factors characterize the most recent wave of modern evaluation protocols, higher specificity of data –meaning more data points observed and rated –and a more generalized, universal set of teaching behaviors. By increasing the specificity of the observed data, teachers are given concrete ways in which they may grow their practice –always assuming that feedback is provided to them in equally specific and constructive ways. Haefele (1993) delivered a scathing review of the state of evaluation, pointing out the flaws in many then-current observation models that provide little complete evidence toward helping teachers in formative –or summative – manner. More recently, observation protocols have been developed that attempt to map teaching behaviors to research-based activities that provide more specific grounds for evaluators to rate teacher performance (see, e.g., Danielson, 2011; Marzano, 2007; Stronge, 2012; Williams, 2009). Of import to the systems used in this study, the

observation protocols used in the districts participating the in the teacher focus groups used for analysis, Danielson's *Framework for Teaching* (Danielson, 2011) and McREL's *Teacher Evaluation Framework* (Williams, 2009) –Danielson and McREL, respectively, attack the issue of specificity of data in similar ways. The Danielson framework evaluates teaching across 22 components clustered around four domains of teaching (The Danielson Group, 2011). Of these components, both 2 and 3 are considered observable during a teacher's normal classroom practice. Domains 1 and 4 evaluate other aspects of teaching that affect the classroom but are not primarily practiced during an observed lesson. Teachers are assessed on each component on a four-point scale of Unsatisfactory, Basic, Proficient, and Distinguished. In total, with 22 components broken into four levels of achievement, teachers are provided 88 total data points to understand for their observation. In practice, naturally, teachers are given fewer to manage, as teachers will likely emphasize the level at which they were rated and the next higher or lower.

McREL's protocol is similar, with 25 elements of teaching divided across five "Professional Teaching Standards" (Williams, 2009, p. 3). McREL also uses a four point scoring scale, which amounts to the potential for 100 data points delivered to teachers. Such specificity, based on research-driven behaviors is a hallmark of protocols being promoted and used in large-scale evaluation reform efforts (e.g (Bill and Melinda Gates Foundation, 2013; Curtis, 2012; Milanowski, 2004; New Jersey Educator Effectiveness Task Force, 2011; Sartain, Stoelinga, & Krone, 2010).

Validity and reliability. Darling-Hammond, Wise, and Pease (1983) recognized the issues of reliability and validity as a major weakness of evaluation methods three decades ago; these criticisms remain today (Bell, Gitomer, McCaffrey, Hamre, Pianata, &

Qi, 2012; Braun, 2005; Toch & Rothman, 2008; Wayne & Youngs, 2003). In general, there is "consensus that current teacher evaluation systems often do little to help teachers improve…" (Darling-Hammond, Amrein-Beardsley, Haertel, & Rothstein, 2011, p. 2). Without valid measures of teacher evaluation, there is little reason for teachers to use the data provided to them.

At present, some scholars are acknowledging the need for a deeper level of combining understanding of the classroom context with evaluation. This focus on valid, reliable measurement combined with an eye toward contextual factors has been dubbed "teaching quality," which separates itself from the more typical "teacher quality" of current systems. As defined by Bell and colleagues (2012), "Teaching quality refers to the quality of interactions between students and teachers; teacher quality refers to the quality of those aspects of interactions that can be attributed to the teacher" (pp. 63-64). Though observational validity is a desirable feature of teacher evaluation systems, other methods of increasing evaluation validity have emerged. Other research has shown that longer observation sessions increase the validity of the observation (Cronin & Capie, 1986).

To supplement the validty of observational data many scholars and policymakers have advocated for the use of student achievement data to evaluate teacher performace. These have primarily been focused around teachers' value-added to student performance, and, more recently, formulas for calculating student growth. Value-added measures attempt to capture the "unique contribution [a teacher] makes to her students' progress," (Corcoran, 2010, p. 4). This means attempting to statistically remove other factors that contribute to a students' achievement –or lack thereof –and isolating only what the

teacher "added" to the student. Put another way, they "make use of current and historical test scores to estimate a teacher's effect on student achievement growth (Hill, Kapitula, & Umland, 2011, p. 3). These measures may account for a student's family background, socioeconomic status, the mix of peer students in the classroom, and any number of other factors in its statistical model (Baker, Oluwole, & Green, 2013). Value-added measurement has recently faced a growing number of criticisms from research around its use as a measure of teacher effectiveness. (Baker, Oluwole, & Green, 2013; Baker, et al., 2010; Corcoran, 2010). These criticisms have caused some policymakers to search for another student measure of teacher effectiveness.

One alternative is commonly called "student growth percentiles" (SGPs) (Betebenner, 2008, 2009). Though value-added and SGPs are different measures, they are often spoken of similar terms. Unlike the statistical modeling of value-added measures, SGPs analyze student gains and losses to comparable students taking the same exams. Simply put by Baker, Oluwole, and Green (2013),

> Some students have achievement growth on the underlying tests that is greater than the median students, while others have growth from one test to the next that is less. That is, the approach estimates not how much the underlying scores changed, but how much the student moved within the mix of other students taking the same assessments…. (p. 7)

SGPs thus attempt to isolate teacher contribution to students' performance by comparing the growth of students year-to-year. If a teacher has a class of students that shows substantially higher SGP than another teacher, one may surmise that the higher SGPs are attibutable to differences in the teacher.

While there has been less scrutiny of SGPs as an evaluative tool, a growing body of literature suggests that SGPs are inappropriate for use as a summative tool in terms of teacher evaluation, whether for rewards or sanction (Baker & Oluwole, 2013; Ehlert, Koedel, Parsons, & Podgursky, 2012; Goldhaber, Walch, & Gabele, 2014). In spite of such criticism, many states have instituted SGP requirements into their teacher evaluation requirements. This is less directly germane to this study, but serves as a reminder that there are no silver bullets in teacher evaluation and thus one cannot simply dismiss a given technique in favor of another. Though teacher observation has its flaws, it may still be a useful tool in adding to a more complete teacher evaluation.

Tied closely to the impacts of validity in evaluation is its counterpart, reliability. Multiple measures and repeated observations have been shown to produce more reliable ratings for teachers (Bill and Melinda Gates Foundation, 2012; Denner, Miller, Newsome, & Birdsong, 2002). Scholars have criticized traditional evaluation measures for their limited frequency of observations, often as little as a single observation of a teacher in a given year (Brandt, Mathers, Oliva, Brown-Sims, & Hess, 2007). Darling-Hammond, Amrein-Beardsley, Haertel, and Rothstein (2012) state that "Successful [evaluation] systems use multiple classroom observations across the year by expert evaluators looking at multiple sources of data, and they provide timely feedback to the teacher" (p.13). They go on to suggest that teachers in a successful system may be evaluated as many as six times per year; other studies have found gains in reliability using 3-4 observations (Curtis, 2012). Reliable observational data allows teachers to more confidently approach data-use by consistently highlighting potential areas of teacher growth or inquiry.

Teacher Characteristics

Beyond the use of high-quality data and potential effects of systemic factors on teacher data use, teachers themselves may affect how data from evaluation is perceived and used. As individuals participating in the change process, teachers will affect implementation and, ultimately, the overall function of the system (Fullan, 2007).

Understanding of teaching. The history of American education features a wide array of educational philosophies. From the progressive movement popularized near the beginning of the 20th century (e.g. Dewey, 1916) onward into the era of accountability and higher standards of *A Nation at Risk* (National Comission on Excellence in Education, The, 1983), education has been fraught with varying opinions on why education matters and for what reasons it is carried out.

Yet, teachers are often prepared by learning instructional methods with little attention paid to habit of self-reflection (Zeichner & Liston, 2013). There are many ways in which teachers conceptualize their purposes as teachers, whether it be as advocates for social justice, or as cognitive and skill enhancers, or as role models (Zeichner & Liston, 2013). Postman (2011) proposed several different "ends" of education within his primary purpose of creating common culture through narratives including stewardship of the Earth and its resources and America as an ongoing experiment and continuous argument. These examples only illustrate the myriad philosophical bents that teachers may possess. Since teachers understand the purposes of education in a variety of ways, the lenses through which they perceive evaluation data may differ, as well. A teacher who sees her primary purpose as working toward developing students who are able to be productive members of a democratic society may interpret scores differently than one who sees his

vocation as a call for disrupting patterns of social or racial stratification. Teachers with differing understandings of teaching likely work alongside one another and are measured by the same evaluation tools, but with potentially different outcomes based on their understanding of how such data applies to their practice. Coupled with new, more extensive evaluation instruments and criteria, understanding of teaching is a potential influencing factor for evaluation data use.

Comfort with evaluation instruments. Teachers in enhanced evaluation systems are likely learning new protocols for observation, as well as perhaps looking at several other new evaluation data sources that may be used in a multiple measures-based evaluation system (Bill and Melinda Gates Foundation, 2010; Sartain, Stoelinga, & Krone, 2010). Teachers' comfort with the instruments of evaluation may play a role in how data use occurs. For example, Datnow, Park, and Kennedy-Lewis (2012) found that teachers using student test data for reflection on their own practice both appreciated that chance to look more closely at student data but also were concerned with the use of testing as potentially overwhelming to students, thus negating some of the effectiveness of data use. This may imply that evaluation systems with multiple measures could push teachers away from data use as they work more to understand the new tasks asked of them instead of applying data to developing their practice. Even in places where teachers are familiar with the process, such as observation, changes to the protocol or who is conducting the observation may influence teachers' comfort. Since teachers already feel at risk from unknown factors in their classroom (Conley & Glasman, 2008), new evaluation tools, and the initial discomfort felt by teachers in adapting to new requirements may play a role in affecting data use. Many scholars have advocated for

necessity of training for teachers and raters in these new systems (Cruickshank & Haefele, 2001; Oliva, Mathers, & Laine, 2009), as well as attention to the specific context of teachers' work (Stronge & Tucker, 1999). Many new observation protocols require teachers and evaluators alike to be trained to encourage effective use of the protocol (The Danielson Group, 2011; Williams, 2009)

Understanding of elements of evaluation. One aspect of developing comfort with the instruments of evaluation is whether or not teachers understand the specific criteria on which they will be evaluated. As noted earlier, some systems for observation may include upwards of 100 specific performance criteria across several facets of teaching performance (Williams, 2009). This may affect how teachers approach an observation session by encouraging them to put on a "showcase lesson" (Nolan Jr. & Hoover, 2004, p. 12). Teachers will adapt their lessons when observed to accommodate perceived desires of the system and the rater. In an example given by Nolan Jr. and Hoover (2011), a foreign language teacher extended a verbal "warm-up" exercise from the typical five minutes devoted in a normal day to twenty-five minutes during observation since the teacher believed the rater was keen on seeing student interaction in the classroom. These sorts of actions suggest that teacher perceptions about the evaluation criteria may play a role in how teachers behave and thus how they may use evaluation data given to them. Nolan Jr. and Hoover go on to suggest that "staff development must be provided for all participants so that each individual has the opportunity to understand the standards and criteria" (p. 17). Other scholars continue to advocate for effective training of teachers in terms of understanding the criteria of their evaluation (Danielson, 2011; Montgomery,

2013). Teacher understanding of the evaluation criteria may then play a role in how teachers perceive and use evaluation data.

Thus there are many potential contributing factors that may play a role in determining how teachers approach, perceive, and use data provided them in their observation sessions. These factors set the stage for this study by suggesting a hypothetical framework by which teacher data use is influenced by these factors. This study will explore teachers' reports about data and data use in new, enhanced teacher evaluation programs in a teacher evaluation pilot program. Using teacher focus groups and survey data, this study explores the ways in which the factors that may affect teacher data use, suggested by the extant research literature, may contribute to or hinder teacher data use. These issues may be seen as creating a type of "feedback filter" that may affect the use of the data and, ultimately, the outcomes of the evaluation program for teachers. While this study is exploratory rather than predictive, teachers perceptions about these issues may help to generate hypothesis about data use processes ripe for testing in further research.

Summary

There is little known at present about the process of teacher data use in evaluation, but a great deal of mounting data to help explore the issue. As teacher evaluation programs evolve and implementation is ramped up under increasing pressure to provide metrics of teacher performance, the question of how these programs work to improve teaching –or whether they work at all –will be a topic of concern. Current research in the field of teacher data use and teacher evaluation have only just begun to coalesce to explore the relationship between the two. While the current body of research

offers several factors that influence educational change and teacher performance, research in data use has not been thoroughly applied to evaluation.

The research conducted for this work explores how some of these factors that affect data use are perceived by teachers –specifically system functionality and data quality –as affecting their ability to use data. Seeing these factors through the eyes of teachers may offer a way to conceive of how these factors may help or hinder teacher data use and to suggest ways to mediate implementation to make the process and practice of evaluation more effective for improving teaching and student learning.

CHAPTER 3: RESEARCH METHODS

Data Mine: Structuring Analysis of Teacher Data Use Perceptions

This research into teachers' use of evaluation data and the factors they perceive as affecting that data use is a secondary analysis of data generated as part of a larger assessment of a teacher evaluation pilot program. As noted above, this analysis combines the use of survey data from teachers participating in the two-year pilot program as well as focus group interviews with teachers in participant districts. Thus survey and qualitative analysis are used to produce exploratory findings regarding the stated research questions, focused on the use of new, enhanced observation protocols. This chapter outines the research methodology used in developing these findings. This includes a description of the overall assessment project in which this work nests, followed by a brief exploration of the contents of the observation protocols adopted by the pilot districts. Finally, this chapter delineates the specific sample, data collection and avenues of inquiry into the research completed herein.

The Teacher Evaluation Pilot Program Assessment

The research outlined here was conducted within a mixed-methodological assessment of a two-year pilot program designed to test a new teacher evaluation system within a single state in the northeastern United States. The program was designed to test potential new regulations for a teacher evaluation program that would eventually be put in place to affect all public schools in the state. The pilot program provided funding and support to ten school districts to select, train on, and implement a teacher evaluation program across multiple measures, of which teacher observation was a major facet. In

Year 2 of the pilot program, an additional fifteen school districts were chosen for inclusion in pilot; each of the original ten districts continued their involvement in the program. Program guidelines allowed districts to choose any robust observational systems within a set of specific guidelines, and provided districts with four preferred protocols: Enhancing Professional Practice: A Framework for Teaching (The Danielson Group, 2011),The Art and Science of Teaching: A Comprehensive Framework for Effective Instruction (Marzano, 2007), The Stronge Teacher Effectiveness Performance Evaluation System (Stronge, 2012) and McREL's Teacher Evaluation Framework (Williams, 2009). All pilot districts selected a protocol from the preferred list, with one district choosing Stronge, two choosing Marzano, three choosing McREL, and the remainder choosing Danielson. These programs are briefly discussed below.

Observation Protocols

In each of the pilot districts a new observation protocol was chosen. All four systems are nationally available and designed to facilitate the observation of effective teaching as well as an understanding of what effective teaching means. Each of the four selected protocols is briefly described, below, with somewhat more focus placed on the Danielson and McREL frameworks, since they were selected by more districts than Marzano or Stronge.

Danielson. Charlotte Danielson's *Enhancing Professional Practice* was first published in 1996, and was "built on the research compiled by ETS in its development of Praxis III: Classroom Performance Assessments, an observation-based evaluation of first-year teachers that is used for the purpose of licensing. The Framework extended this work (examining current research) to capture the skills of teaching required not only by

novice teachers but by experienced practitioners as well" (Danielson, 2011, p. 1). The framework was revised in 2007 and again in 2011 in response to the MET study (Bill and Melinda Gates Foundation, 2010; 2012). The 2011 revision did not have substantive changes to the Framework, but revised language and provided exemplars for use with training observers for the MET study.

According to the Danielson Group, "The Framework may be used for many purposes, but its full value is realized as the foundation for professional conversations among practitioners as they seek to enhance their skill in the complex task of teaching" (2011, p. 1). The Framework is thus not specifically designed for personnel decisions such as merit pay, hiring and firing, or tenure; it is meant as a professional development tool to allow teachers to develop their skills to become more effective.

The Danielson framework evaluates teaching across 22 components clustered around four domains of teaching (The Danielson Group, 2011). The four domains are: 1. Planning and Preparation, 2. Classroom Environment, 3. Instruction, and 4. Professional Responsibilities. Each domain contains 5-6 components, each of which represents an aspect of teaching to be evaluated. These can range from "Demonstrating Knowledge of Content and Pedagogy," Component 1a; to "Engaging Students in Learning," Component 3c; to "Maintaining Accurate Records," Component 4b. Teachers are assessed on each component on a four-point scale of Unsatisfactory, Basic, Proficient, and Distinguished.

Of these components, both 2 and 3 are considered observable during a teacher's normal classroom practice. Domains 1 and 4 evaluate other aspects of teaching that affect the classroom but are not primarily practiced during an observed lesson. Thus teachers'

perceptions about the observation process in districts using Danielson responded in general about domains 2 and 3.

McREL. Three participant districts selected the McREL protocol. McREL's Teacher Evaluation System (Williams, 2009) is similar to Danielson in many ways, the first of which is its aims. "The intended purpose of McREL's Teacher Evaluation System is to assess the teacher's performance in relation to the Professional Teaching Standards and to guide the creation of a plan for professional growth" (Williams, 2009, p. 1). Like Danielson, McREL lists its purposes as formative for teachers, with goals of helping teachers reflect on their own practice and districts to design professional development programs for those teachers. McREL does not state outright that the tool is intended for summative purposes.

The evaluation is divided into five "Professional Teaching Standards:"

I. Teachers Demonstrate Leadership
II. Teachers Establish a Respectful Environment for a Diverse Population of Students
III. Teachers Know the Content They Teach
IV. Teachers Facilitate Learning for Their Students
V. Teachers Reflect on Their Practice (Williams, 2009, p. 3)

Each standard is divided into a number of elements. The number of elements in a given standard ranges from three in standard V to eight in standard IV, with 25 elements in all. Like Danielson, teachers in McREL are rated on a four-point scale: Developing, Proficient, Accomplished, and Distinguished. To achieve a given rating in an element, teachers must fulfill the requirements of that rating and all lower ratings. For example, to

receive a rating of "Accomplished" for element 2c, "Teachers treat students as individuals," a teacher "holds high expectations of students **and** communicates high expectations for all students **and** encourages and values contributions of students, regardless of background or ability" (Williams, 2009, p. 15).

The McREL protocol focuses on interaction between the evaluator and teacher. The system calls for the teacher to complete a self-assessment in which they score themselves in each of the elements, as well as a pre- and post-conference about the observed lesson. Though Danielson has the pre- and post-conference discussion features, they are not highlighted in the framework text itself as in McREL.

Marzano. Two districts in the study selected *The Art and Science of Teaching: A Comprehensive Framework for Effective Instruction* (Marzano, 2007) as their new observation protocol. Like the others, the Marzano framework is aimed at providing feedback for teachers in a formative sense and is built for teachers to reflect on their own practice as well as be rated by observers. The framework is divided into four domains of teaching. Domain 1, Classroom Strategies and Behaviors, is the observed domain, and is sub-divided into 41 key teaching strategies across differing segments of the observed lesson. The other domains deal with planning and preparing, reflecting on teaching, and collegiality and professionalism.

Stronge. The Stronge Teacher Effectiveness Performance Evaluation System (Stronge, 2012) is perhaps the most attuned to the content and contextual specifics of schools using the tool of the four protocols. The system revolves around seven "Performance Standards" (Stronge, 2012, p. 4) that are elaborated by a number of

performance indicators. Like the other programs, teachers are rated in each in one of four categories, ranging here from "unacceptable" to "exemplary" (p. 9).

Each of the observation protocols shares a good deal of similarity and addresses common teaching behaviors with an eye for formative assessment of teachers' practice. Understanding the basics of these protocols sets the stage for interpreting teacher perceptions and potential use of data.

Additional Measures

Schools in the pilot also began to implement measures of evaluation based on student growth percentiles (SGP) for students in math and language arts in grades 4-8 and student growth objectives (SGO) for all other grade levels and courses (Betebenner, 2008; Betebenner, 2009). Though there is considerable debate around the use of student growth scores to evaluate teachers (Baker & Oluwole, 2013; Baker, Oluwole, & Green, 2013; Baker, et al., 2010), the pilot program elected to include these measures in its overall evaluation program. SGP uses state test data –drawn from the statewide standardized test – to show students' change in score from year-to-year in the exam. The students are then organized by class and the mean growth for a given teacher is compared with that of other peer groups of students to measure teacher effectiveness.

SGOs are typically teacher-developed pre- and post-tests tailored to specific course content. Students' growth here is determined by the outcomes of the two exams and compared against a target determined by the teacher, school, or district. The student performance measures were in the nascent stages of implementation during the year of

data collection for this work; and though more present in Year 2 of the pilot, they were neither well developed enough for, nor particularly germane to, this work.

The assessment study included multiple measures designed to capture districts' progress in implementing the new evaluation program. These included surveys of teachers and administrators in all ten pilot districts –a separate survey was administered to teachers and two surveys were given to administrators –and site visits to six participant districts of the ten. Site visits, the source of all qualitative data analyzed for this work, included interviews with district administrators and focus groups with teachers. In the second year of the program, teachers and administrators were surveyed again –only one survey was given to administrators in Year 2 –and site visits were conducted in 10 districts. Four of these were conducted in districts previously visited in Year 1, and six were to districts new to the pilot in Year 2.

District Sample

The districts involved in this study were diverse in a number of ways. Participating districts were geographically varied, and were comprised of schools in urban, suburban, and rural settings. Additionally, the size of the districts varied greatly, with the smallest districts having enrollments around 1,000 to the largest having enrollments of nearly 24,000, with most districts serving between 2,000-5,000 students.

In addition to their differences in size, pilot school districts were diverse in the socio-economic makeup of their students. In the pilot state, school districts are classified economically by the use of a specific formula, ranking districts by income levels. School districts participating in the pilot ranged from the lowest socioeconomic bracket to the

highest according to the 2000 census data, the most recent calculation available for this information. Though not selected purposefully to create a statistical sample, the income levels of participant districts encompassed the range from poorest to richest across the state.

The districts in the pilot cannot be considered representative of the state as a whole for a number of reasons. Districts self-selected themselves as grant applicants prior to participation, and as such there were only 10 districts in Year 1, and 25 in Year 2, from which to draw a sample. However, the diversity of locations, socio-economic makeup, and size gives a broad palette of experiences for teachers and shares characteristics with a large proportion of districts across the state.

Research Methodology

As previously stated, this work uses analysis of selected data collected during the assessment of the teacher evaluation pilot program. Specifically, both the Year 1 and Year 2 teacher survey and the focus group interviews from site visits in both years were used to analyze the research questions. The specific methodology for this research is outlined below. This includes a brief description of the observation protocols selected by districts as well as specific description of the research tools, collection of data, and analytical techniques.

Sample

The sample for this research is drawn from two groups. The first is the larger group of participant teachers, each of whom were invited to participate in a survey during

each year of the assessment. The second is the sample of teachers interviewed during site visits to pilot districts.

Surveys. In the spring of 2012, teachers from all ten Year 1 pilot districts were asked to participate in a 40-question survey encompassing several aspects of their experiences in the pilot. 2,495 responses were collected from 4,229 invitations to participate, a 59% response rate. Response rates from individual districts ranged from 36% to 100%. Survey data was collected and analyzed using Qualtrics (2013) software. The Year 1 Teacher survey is shown in Appendix A

In Year 2 of the pilot, teachers now from all twenty-five participant districts were again asked to respond to a survey about their perceptions of the pilot program. Response to this survey was more limited; the response rate for the Year 2 survey was 39% with 2926 responses collected. Individual responses rates from the districts ranged from 11% to 100%. The Year 2 survey is included in Appendix B.

The survey, in this case, proved to be a tool to search for potential perception questions which could then be framed by interview analysis. This allowed the research to pursue topics that were not specifically targeted in the interviews. Combining teacher survey data with more detailed interview analysis allowed for a more focused construction of the conceptual model, as survey questions could be used as indicators of potential affective factors for data use that could then be expanded upon with interview data. Conversely, topics that teachers mentioned frequently or infrequently in focus groups could then be searched for in the survey to assess how specific information provided in the focus groups compared to the impersonal survey responses.

Teacher Focus Groups. From each of six of the Year 1 pilot districts, two teacher focus groups were interviewed. Focus groups ranged in size from 4-8 teachers depending on the number scheduled by the district. Districts were asked to provide 6-8 teachers for the focus groups, mixed between secondary and elementary levels. Selection of teachers was left primarily to the pilot project director in each district. Directors were asked to provide a mix of tenured and untenured teachers as well as teachers from tested and untested subjects –in this state, tested subjects are Language Arts and Math in grades 4-8 and 11. Though the original design sought to hold one focus group of teachers of tested subjects and one of teachers of untested areas, teachers were generally mixed in both groups in a given district. In one district –two focus groups –only middle- and high school teachers were available; another district did not have a high school –they sent students graduating from their middle school to a regional high school –and thus only elementary and middle school teachers were interviewed. The focus groups contained a mixture of men and women and teachers of varying levels of teaching experience and content specialties.

During Year 2 of the study, ten districts were visited and two focus groups conducted in each. The number of teachers interviewed in these groups was consistent with Year 1. Rather than mixing teachers of all levels simultaneously, however, Year 2 focus groups with conducted as one elementary level group and one from the secondary level. The secondary focus group contained a mix of grade 6-12 teachers depending on those chosen by the project director and grade levels served in the district. Four of the site visits in this year were repeat visits to Year 1 sites; the remaining six were to districts new to the pilot district in Year 2. In total, twenty additional focus group interviews were

completed, two in each Year 2 district. Like Year 1, focus group participants had a wide range of experience teaching, subject areas, and were varied with regard to teacher race and gender.

That these heterogeneities among the focus groups existed –that is, that there were few clear breakdowns present in each of the groups as to experience, discipline, gender or grade level –should have little bearing on the collected data. All teachers in the pilot were to experience the same observational process, though tenured teachers received fewer total observations than non-tenured teachers, as per the pilot regulations. On the whole, the mix of teachers in the groups likely allowed a comparison of experiences that highlighted potential differences in the process among reasonably disparate groups of teachers within each district, as teachers could compare their observation to colleagues who had been observed by both similar and different raters across grade levels, schools, and years of experience.

Data Collection

Site visits for Year 1 focus groups conducted during the spring of 2012; Year 2 focus groups were conducted from Fall 2012 through Spring 2013. Participants were asked to sit for a 45 minute focus group style interview. Actual focus groups lasted from 35-55 minutes. An interview guide (Patton, 1990) was developed to guide discussion while giving the interviewer freedom to develop rapport with participants and pursue interesting topics of conversation. In Year 2, a site visit protocol (Yin, 2008) was added to the procedure to aid in collection of data. All focus groups were audio-recorded. The interview guide is included in Appendix C. The site visit protocol is found in Appendix D.

Among other focus group topics, teachers were asked to describe qualities of an observer that could best evaluate them, their impressions about the usefulness of the observation protocol and feedback derived from it, and what parts of the evaluation pilot were going better or worse than previous systems. The focus groups and interviews were audio-recorded and transcribed for analysis in Year 1. In Year 2, focus group data were compiled in a site visit guide for analysis.

Data Analysis

Survey Analysis. Survey data was primarily used descriptively, as indicators of teachers' agreement with certain statements or claims made about their behaviors were used as markers to suggest themes and potential avenues of research for the qualitative analysis. Oftentimes, these questions helped to set or confirm the initial directional aspect of the hypothesis for a given affective factor being explored –whether a given factor was seen by teachers as helpful or harmful to their ability to use observational data. One example of this procedure is the exploration of time as a factor that affects teachers' data use. The original hypothesis for this factor as suggested by the literature (e.g. Hall & Hord, 2006; Huberman, 1983) was that teachers would feel that if the evaluation pilot was perceived to take too much time, teachers would feel it constrained their ability to use data. In the Year 1 survey, teachers were asked whether they felt the program took too much time. The survey responses showed that a large majority –around 65% -of teachers felt that the program, indeed, took too much of their time. This result was then applied two ways. First, as a confirmation of the findings of the extant literature –teachers did find their time constrained by this change. Second, since the finding pointed toward

analysis of the qualitative data for time, a code for time-related comments was developed and these comments subsequently explored.

Simple cross-tabulation was employed in some cases in order to examine the responses across multiple questions. An example of how cross-tabulation was applied in this study was in trying to determine teachers' forms of data use. Teachers reporting that they used data to improve their practice or to gather higher scores on future evaluations were directly asked these questions, and their responses were measured directly. As focus group interviews suggested that some teachers were not using the data at all, responses to these two questions of data use were cross-tabulated to uncover those responses that indicated no data use. These data are shown below, in Table 5. Using the cross tabulation to find teachers who did not report any observational data use served as a platform to frame the additional analysis. Reasons for data non-use could then be explored in the context of the number of teachers who reported such use, giving breadth to the deeper, narrower, focus group interview analysis.

While the survey data often showed potential avenues for further analysis, it was also used as a confirmation or refutation of the qualitative analysis. When few teachers mentioned using data for improving their teaching in the focus groups, the survey data was consulted to look for similar issues. Here, the survey data was somewhat more positive than qualitative themes, as shown in Table 3. This then led to a more balanced analysis, highlighting relative teacher positivity in the survey, tempered by focus group indicators.

Focus Group Analysis. The qualitative analysis draws upon the work of Miles and Huberman (1994) and Yin (2008), following an iterative process that begins with

literature-based conjecture, successively assesses those conjectures against the observational data, and refines them to more adequately reflect those data. As discussed above, some examples of literature based conjecture used dealt with the issue of time as a constraint, as well as the initial hypotheses that teachers would not trust external observers, drawn from the work of Conley and Glasman (2008) and Yuan and colleagues (2013), among others. Using the literature to guide the intitial delineation of potential affective factors allowed for the construction of the conceptual framework, supported by pillars of existing knowledge on the subject.

In addition to codes related to the factors suggested in the literature, data analysis was conducted both deductively and inductively. Original parent codes were inductive, based on questions from the interview guide dealing with questions related to the research questions. Examples of this type of inductive coding include the exploration of the usefulness of data –teachers were asked in the interview guide what they thought would make their observations more useful –which was then used in early rounds of coding.

Other parent codes were developed deductively around readily apparent major themes in early analytical rounds such as the existence of specific barriers and facilitators of data use, characteristics of observers, and examples of teacher data use. In the first phase, preliminary ideas and hypotheses regarding the research questions were developed and emerging themes of the data were noted. The initial parent codes for the teacher focus groups were based primarily on hypotheses from the research questions as well as directly from questions. These codes were simple descriptors such as “Content Knowledge” and “Fairness and Reliability.” Additional parent codes were generated from

survey questions and preliminary results. For example, teachers appeared somewhat positive on survey items regarding their perceived improvement of teaching quality based on observational feedback received. One parent code hypothesized from this was "Data Use: Improved Teaching"

Using Dedoose (v. 4.5.95, 2011) software, a qualitative and mixed-methods analysis program, the focus group interview data were coded with regard to teachers' perceptions about evaluation data use and the factors teachers suggested affected this process.

The first rounds of analysis relied on several techniques described by Miles and Huberman (1994). Early rounds focused on *clustering,* used for "grouping, then conceptualizing objects that have similar patterns or characteristics." (Miles & Huberman, 1994, p. 249). Clustering was of primary importance in solidifying hypothesized and emergent parent codes as well as developing child codes. A key early-round use of clustering was the data use comments of teachers. Several teachers mentioned using post-observation feedback as a point of dialog with observers about scores. Clustering of teacher comments in interview transcripts suggested that this was a reasonable parent code, as several statements fell into this overall concept. A rule for coding was developed that comments where teachers noted their discussion of scores, either past of future, as well as learning about the protocol would be included in this cluster. Examples included:

> It is time consuming, the post-conference is you know? It takes a long time to go through it all and go through every single [component]… I had an instance where, if I'd said, 'Okay, do you have a[n American] flag in

> the room?' They missed that. They didn't see it, but I had American flag in the room you know, I mean it's not…and you go through every single thing.
>
> I think the dialogue you have at post observation is important because I know in my first [observation] something was suggested and we had done it like a couple weeks ago. It pulls some things together or we were waiting for another time to do it at the end of the week because maybe it was half day. So [the observers] accept that. It's good to hear the suggestion when you know you either have done it, or you are getting ready to do it, they just didn't see it there.
>
> The classroom observation tool, I think it's very lacking saying as far as being at the extreme end. I don't think that it gives you… it doesn't delineate exactly what the teachers need to do or like some exemplars or it's not explicit enough.

These types of comments, teachers talking about discussing their scores with observers, concerned about how to succeed within the framework, or working to improve scores via rebuttal, compliance with new protocols, or seeking loopholes, were all connected by the common characteristic of learning about the new system. Other excerpts related to this cluster were then specifically sought in future iterations. This added to the weight of the overall cluster as a meaningful finding. In this case, data use to learn about the system was then looked for in the survey data, and the two data sources combined framed the

findings presented below related to teacher data-use to defend current ratings or increase future scores.

Further, this clustering suggested further breakdowns into child codes. As successive iterations of the analysis employed clustering, subtleties in the qualities of data use for learning responses arose. Rather than teachers merely trying to learn about the system, sub-clusters were noted that suggested two distinct ways that teachers were interacting with observational data, using their learning in slightly different ways. These led to the formation of child codes around data use for rebutting/defending ratings as well as for ensuring higher scores in future evaluations. When teachers reported clarifying scores with observers, or, as in the quote above regarding the flag, pointing out behaviors that were not accounted for in the current score, these comments were clustered around the child code dealing with rebuttal or defense of current scores.

Other teacher comments were oriented toward the future, such as "It's putting us, it's putting teachers, I think, what you said, telling us we are getting to the point where it's what do you want us to do we'll do it…." These type of forward-reaching comments became a second sub-cluster, and further comments referring to data use for future observations were then coded into this cluster. While still related to the overall idea of learning the system, these comments tended toward looking to improve scores on future evaluations rather than modifying current scores. Clustering these comments together and reviewing them over several iterations allowed these subtleties to be addressed.

Another analytical technique employed was making metaphors. Metaphor simultaneously allows further pattern-making while reducing data noise without eliminating avenues of inquiry (Miles & Huberman, 1994). One notable place that this

technique was used in the analysis was also within the data use for learning about the system cluster, dealing with achieving better scores in the future. That metaphor was "gaming the system." In this case, the metaphor was supplied directly by the teacher in a focus groups comment describing how he was using the data. When this metaphor was noted in the Year 1 focus group interviews, it spurred the recollection of the same metaphor being used in a different district's Year 2 secondary teacher focus groups. This colorful metaphor, used in two districts across two years, prompted the creation of a sub-cluster of learning about the system.

These types of comments were separated from compliance by the active use of the protocol and observational data to achieve a higher rating. Compliance comments expressed interest in simply having clear understandings of the various levels, a "tell us what we need to do" mentality; "gaming the system" was related to using this knowledge to achieve a higher score. This became a basis of separating future-facing data-use comments.

"Gaming the system" allowed for further depth of analysis, as the survey findings regarding teachers who had changed their teaching to receive higher scores were now colored in a broader spectrum. Some teachers were changing practice for higher scores as a "go along to get along" –a metaphor itself –action, while others sought, or, at least, considered ways to gain higher scores directly for their own benefit. Though the analytical metaphors used here were not so prevalent as to be able to generalize to all teachers –this is not the goal of qualitative analysis –it provides a depth of understanding for the reasons behind teachers changing their behaviors that was not possible in the survey data.

For factors affecting data use, metaphors were used in later iterations, when determining how to make sense of code groups that expressed concerns in a multitude of ways and depth of severity. One informally applied to teacher statements on the purposes of the new evaluation system was along the lines of "take us down a peg." Using this metaphor highlighted teacher concerns about a perceived threat in the summative use of evaluation data. The process in the metaphor creation and subsequent coding were similar. With the creation of the metaphor –often drawn from teacher comments themselves –came a basis for coding; this coding cluster was then compared to survey data and other clusters to assess how important it was in to explaining the various research questions and affective factors.

Other qualitative techniques were also used. Since teacher focus groups incorporated many teachers across 32 focus groups in 11 different districts, some across two years of data collection, a number of opposing viewpoints surfaced. In these situations, counting and making comparisons was used to help determine the weight of the evidence. These simple qualitative techniques were used to analyze the overall sentiment of the larger group of teachers, while allowing contrasting or dissenting voices to be heard. One example of this came through in teachers' perceptions about the perceived necessity of observer content knowledge and the use of external observers. By simple count, using the "positive" and "negative" comments around whether content knowledge was necessary in observers gave strong evidence of teacher preference for an observer with content knowledge. Simultaneously, counting showed a strong negative preference for external observers. Counting was used more frequently in Year 2, as the

nature of the site visit guide made looking for similarities and differences among various focus groups and districts a simple yet useful tool for analysis.

Yet when comparing comments about these two issues, a question arose. Since teachers prefer content-knowledgeable observers, and since access to content specialists in all subjects is unlikely within the administrative ranks of a given district, was teachers' dislike of external observation affected by this conundrum? This led to the possibility of an additional theme: potential upsides of external observation. This comparison spurred analysis in a new direction, both widening the search items as well as providing depth to the issue of external observation.

These techniques were applied across cases and through various iterations of reading and coding the data. Triangulation occurred primarily through similar sentiments arising in different districts. Additionally, opinions about the program demonstrated a range of acceptance, helping to provide a more balanced analysis, since the overall sentiment of participants was neither wholly positive nor extremely negative.

Year 2 site visit guides were analyzed in largely the same manner, though there was less reliance on creating new codes; there were fewer direct quotes. Most themes, therefore, were developed during the coding process of Year 1 participants. Year 2 analyses helped to color and confirm those themes.

Summary

This study used the data collected from a larger assessment of a teacher evaluation pilot program in one state to explore the research questions outlined above. This included a descriptive analysis of surveys completed by teachers in each of the two years of the pilot program. These survey findings were used to both provide direction for

analysis of the qualitative data related to factors that affect teacher data-use as well as to provide confirmation or refutation of qualitative themes. Several description survey analysis techniques were used including cross-tabulation.

The more in-depth qualitative analysis was based around multiple coding iterations of the focus group interview transcripts from Year 1 and the site visit guides created from Year 2 interviews. This coding process employed a variety of classic qualitative techniques including simple counting, creating comparisons, clustering, and metaphor making. Beginning from literature-based hypothesis making, qualitative coding was conducted over multiple readings of the dataset, and themes induced from these findings were noted in the findings.

The use of qualitative data adds depth to the findings of the survey data. The survey data, in turn, provides a larger lens through which to view more individual themes present in the qualitative aspects of the research. As this study is exploratory in nature, the use of both survey and focus group data provides a well-rounded base of support to the conceptual framework, suggesting both larger and finer points to aid in fleshing out the conceptual framework.

CHAPTER 4: FINDINGS

Same Old Thing: Teachers' Perceptions of Data Use from Enhanced Observation Protocols

Over the course of the pilot program, teachers experienced a variety of interactions with the observation protocols chosen by their districts. The new protocols were designed to give teachers specific, concrete feedback to promote formative development of teaching practice. Pilot districts used these protocols in multiple ways, aiming to encompass both the formative aspects of teacher observation as well as creating summative measures to describe overall teacher performance on an individual level. To comply with the development of an overall summative rating for teachers across both observation and student growth measures, observation component ratings were assigned numerical values, i.e. a "Distinguished" rating was marked as a "4" while "Proficient" was counted as a "3." The observation protocols in their published formats did not include numerical rankings for each given assessment level. Essentially, any observed behavior from the protocol was assigned both its text rating and its corresponding numerical value. These ratings served to increase the amount of data available to teachers for consumption. Not only were teachers able to use the formative ratings from observation, but also a summative score for either individual elements of the protocol or for overall rated performance. Additionally, many observers included written prose defining the observer's perceptions of what happened during the session. Thus, in many cases, teachers were provided opportunities to conceptualize their observational data in multiple ways.

This chapter explores teachers' perceptions about how they used data provided to them from observations. In the Year 2 survey, teachers most often reported that they received data from the evaluation process in one of two ways. First, both the selected observation protocols and the pilot program regulations required that teachers receive a post-observation conference with their rater. This gave teachers and raters the opportunity to connect face-to-face to discuss the results of the observation. Second, some teachers received data electronically, as data management systems used by participant districts notified teachers of the availability of their evaluation results and sometimes provided a forum for further analysis, discussion, or rebuttal of those results by the teacher. This, too, was a requirement of participation in the pilot. These methods of data transfer set the stage for how teachers' saw themselves interacting and using –or not using –the observational data.

This chapter will explore teachers' perceptions about the overall usefulness of data provided to them from observation sessions, and the ways in which teachers did –or did not –use data. Overall, some teachers reported that they had made adjustments to their practice to either improve their instruction, to improve their observation scores, or chose not to use the data they received. Each of these effects plays a role in how teachers are affected by other factors that influence their data use.

Perceived Usefulness of Data

The most straightforward question to explore was whether or not teachers felt that the new observation and evaluation systems provided useful feedback. There are two ways that teachers may use observation data: either to improve their practice or to get a better score on their evaluation as an end unto itself. Additionally, teachers may find the

data of little to no use, and elect not to use the data at all. During the teacher survey, teachers were asked several questions to help to understand their feelings about the usefulness of the data. In Year 1 of the pilot program, teachers were asked to compare the usefulness for providing guidance to teachers in the newly implemented system to their previous evaluation system. Table 1 shows the results of this question.

More teachers in Year one felt that the new protocols being used in their districts were better at providing guidance to teachers that felt the new systems were worse. In total, 36.86% of respondents felt that the new system was at least better than the previous system while 21.97% felt that the new system did a worse job of providing useful

Table 1

Perceived usefulness of the observation system for providing guidance to teachers, Year 1 (n=2477).

Much Better than previous system	Better than previous system	Neither better nor worse than previous	Worse than previous system	Much worse than previous system[1]
7.55%	29.31%	31.49%	11.47%	10.5%

guidance. To read the data in a more negative light, one could say that *only* 36.86% of teachers found the new system to be better than their previous one in terms of guidance. Whether this constitutes a failure on the part of the system is a matter of opinion. For the purposes of this study, it is accepted that teachers will display a wide range of perceptions about the new system, and thus this shows a reasonable balance between useful and not useful perceptions, slightly shifted toward teachers finding the data useful.

[1] Other possible answers were "Does not Apply" (1.45%) and "Don't Know" (8.24%).

Additionally, Year One teachers were asked directly whether they felt that the feedback they received from observers was useful. They were also asked more broadly whether the new system provided constructive feedback for individuals and promoted professional development. The responses to this question are shown in Table 2. Once again, more teachers felt that the feedback they received was useful, and that the evaluation system promoted individual and corporate development than those who did not. This indicates that at least some teachers are seeing an improvement in the system to give them an opportunity to use data, but what remains to explore is just how teachers perceive themselves using the information derived from the new system. Here, the balance is shifted more toward the side of usefulness, with no more than 20.1% of respondents disagreeing with the usefulness of the feedback provided.

Table 2

Usefulness of feedback provided under the new observation system, Year 1 (n=2319)

	Strongly Agree	Agree	Neither agree nor disagree	Disagree	Strongly Disagree
I receive useful feedback from observers	6.38%	39.72%	35.92%	12.81%	5.17%
The system for assessing teachers generates assessments that provide constructive individual feedback and promote professional development	5.95%	44.42%	24.67%	12.98%	7.12%[2]

[2] Other possible answers to this question were "Does not Apply" (0.26%) and "Don't Know" (4.61%)

Digging deeper, focus groups during both years of the pilot program were asked whether they found the information provided to them as feedback to be useful. Here, the results were more mixed. In each district, at least some teachers reported receiving feedback they found useful. Some teachers appreciated the opportunity the post-observation conferences provided to engage with evaluators. Said one teacher, "I liked that it opened more dialogue." A second was in agreement. She said, "It was nice to be able to sit down and you know have input on you know what occurred in your lesson." Another responded to these comments with some reservations in terms of usefulness, saying, "Yes, you're able to discuss what went on in the lesson and then I don't… I don't know that if any changes occurred. Like, you had a conversation with your principal; did a change occur on the status of [the evaluation]?" This teacher acknowledged that in her district there had been more opportunity for conversation; she was less certain that the discussion had influenced the evaluation.

In a different district, the opportunity for directed feedback was considered helpful. "So you will select something that you would like that observer to come and look at and then you'll get feedback on it saying, 'You're doing it right or you're doing it wrong or why don't you try this and the suggestion for that.'" In other districts, in Year 2 focus groups, teachers noted that the opportunity to target a specific teaching behavior, such as questioning techniques, was useful.

Though in many cases teachers reported that they found feedback –primarily in post-observation conferences and informal discussion with the evaluators –to be generally useful, other teachers reported that they did not find it helpful. In one district,

only a single teacher reported that the feedback she received was useful in any way; other members of the teacher focus groups in that district stated that they found feedback to be mostly useless. As one teacher from this district put it, "That was my biggest frustration. I'll put the time into it if you can help me become a better teacher but there was very little feedback to improve." Teachers in this district felt they had not received useful feedback; this was consistent in this district across both Year 1 and Year 2.

There was ample evidence that some teachers considered the feedback they received to be useful and others that did not. This led to questioning in what ways teachers were using –or ignoring –the data. Understanding the ways in which teachers were or were not using evaluation data may shed some light on how their data use process interacted with other affective factors. Further analysis of survey and focus group data led to three main data use themes coming to light.

Three Types of Data Use

The first of these was what the designers of the observation protocols, policymakers, and many administrators would likely describe as the intended use of the data: teachers using data gained from observations to alter their practice to improve teaching. A second mode involved using the information presented to them to better understand the protocol requirements. This manifested as data use to argue or rebut the scores they received, or changing their practice to garner higher scores on further observation whether or not the teacher felt that the changes were in the best interest of effective teaching. Finally, teachers claimed not to use observational data. Teachers in these instances commonly perceived that the observational data was not useful to

improving, altering, or maintaining their practice. Each of these three facets of data use will be explored in greater detail below.

Data Use to Change Practice

The formative goals of the teacher evaluation pilot program were to help districts, schools, and teachers improve the quality of teaching. By using enhanced observation protocols aligned with research-based measures of effective teaching, teachers were to receive feedback designed to help teachers and raters identify an individual teacher's strengths and areas in need of improvement. Ideally, teachers would use this data to alter their practice in ways designed to more effectively teach students. This is the idealized form of data use, one that conforms to the skeleton framework illustrated in the introductory chapter. It is the theory of action by which policymakers and implementers presume that change will occur in the presence of updated evaluation protocols. Unfortunately, while there were some positive indicators of this sort of data use, particularly in the survey data, speaking with teachers in person during the focus groups tended to reveal that teachers used this method of data use less frequently than others.

The use of data to alter teaching was seen most positively in the surveys. In the Year 2 survey, teachers were asked to evaluate whether feedback received from observations had affected their classrooms and teaching. Results from this question are shown in table 3, below. The question shows that roughly 40% of teachers report an improvement in the learning environment of the classroom or in their instruction based on feedback from their observations. This finding suggests that a significant minority of teachers believe they are using the data in a formatively positive way, growing in their practice due to data use. The converse of this, of course, is that slightly less than 60% of

Table 3

Teacher perceptions about changes in their practice due to observation feedback, Year 2

Feedback from observers helped me improve…

	Agree	Neither Agree Nor Disagree	Disagree
…the learning environment in my classroom (n=2619)	40.17%	33.37%	26.46%
…the quality of my instruction (n=2616)	40.67%	32.42%	26.91%

teachers are *not* reporting improving their instructional methods based on feedback, a type of data use discussed in more depth below. Nevertheless, this is a positive signal that some teachers are using data from observations to improve their teaching –at least to their minds. Indeed, if this were the depths to which this question was analyzed, one could perhaps call the overall process a great success, with a large number of teachers reporting improving their teaching as a direct result of data use in evaluation. The qualitative analysis of this issue, however, suggests that this type of data use is not as widespread as survey responses suggest.

In focus groups, teachers were less positive than the survey as to their perceived change in instructional quality. In only a small number of cases, half of the districts visited in Year 1, did teachers mention they saw data from observations as a chance to improve their teaching. Some comments from teachers were markedly positive in their attitude toward altering practice. In one district, a teacher stated,

> I myself was very open to it, and [the observer] gave good suggestions, so and I put them into practice. As you teach many years, sometimes you forget these little

> things that you used to know when you started out, you know? Why aren't things working so well? You need somebody to look at it and give you some feedback I think.

This teacher welcomed the opportunity to receive observational data and believed that she had used that data to alter her teaching.

Other comments about data use for improving teaching were less concrete. "I think the good thing is that it does promote discussion, it does promote dialogue, it promotes also teacher growth," said a teacher in a second district. In a third district, another teacher had a similar insight, saying, "It would be nice if it was consistent year to year so the person is making suggestions about my improvement would see the improvement the following year." While not directly stating that the data received was being used to inform teaching practices, these statements imply that their teaching would grow in quality based on feedback.

In Year 2, teachers in five of the visited district mentioned that they found the data received useful. In these districts, however, teachers reporting useful data tended to be more isolated than consensus. Most comments about data use for teaching improvement were intermingled with those of teachers who did not find the information received helpful or used the data as a way to understand the evaluation requirements more thoroughly or as a mechanism to achieve a higher score on future evaluations.

Data use for improving the quality of instruction had some positive indicators as identified by teachers, but the frequency of these reports is somewhat lessened more limited number of reports coming from teachers in focus group interviews. While it may be considered a success of the system when teachers change their teaching for better

instruction, there may be diminishing returns in the case where this form of data use – arguably the only intended form of data use for overhauling evaluation procedures in the formative sense –is the least frequently reported type.

Data Use to Defend/Increase Evaluation Scores

Somewhat more common that using data for raising instructional quality was data use of a slightly different type. This can be described as using data as a tool to understand the observation protocol. The new evaluation protocols selected were often completely new to the district, or, when a protocol was chosen due to the district already having experience with the product, was implemented in a more direct or focused way. Thus teachers were learning the system while they were experiencing the pilot program. The results of this type of data use were typically teachers saying that they found the data useful in two ways: in defense or rebuttal of observation scores, and as information to score better on future observations.

Defense or Rebuttal of Ratings

In the Year 2 survey, teachers were asked whether they believed that the new observation protocol allowed teachers to explain their decisions and actions. 44.47% of respondents (n=2,739) either agreed or strongly agreed with this statement. This is roughly the same percentage of teachers that felt that their teaching had improved as a result of feedback received, 40.67%. Though, again, this not an overwhelming number of respondents, it represents a sizable minority expressing this belief. Unlike the use of data for improving the quality of teaching, however, statements from teachers in focus groups were frequently recorded.

"I think post-observation, to have an observer who's willing to hear your feedback or…they give you feedback and they're willing to hear your response to that feedback. I think that is important," said one teacher, who valued the opportunity to give her impressions of the observed lesson. She felt that is was important to have the opportunity to discuss ratings given and use the perceptions of her rater to be able to dialogue regarding performance ratings. Another teacher in the same district, but a different focus group, had used the data to defend her teaching. She said, "I have been questioned why [I did the questioned behavior]. And it didn't end up in my observation because I was able to explain myself." This teacher is not using the data she received as a tool to improve her practice but instead as an indicator of where her teaching may be, in her estimation, under-rated or under attack. In another district, teachers wanted to use the data provided to defend their practice, but were concerned that the new protocol limited their opportunity to do so. "Something that happened at our level, if you received a basic… there was no room for explanation unless you wrote a rebuttal," noted one teacher. When asked for clarification about what alternative there might be to written rebuttal, the teacher replied, "In the past, before we signed off on the observation, there was a dialogue. And in that dialogue if something wasn't clear, the teacher had the opportunity to say, 'Well this is what I did,' and so that could be changed before you sign your observation."

Several other comments, each from a different district, reflected the use of data for defense or rebuttal of ratings:

If you can explain what you are doing and they are going to listen to your explanation and listen you know properly and ask you some critical questions then they should be able to personalize it

Leader: Did you move from a three to a four [rating score] for instance?
Female Speaker: On one particular subtopic I had done, yes.
Leader: Okay.
Female Speaker: Because I did express something and he goes, "oh you're right I didn't look at it that way," so…

It is time consuming, the post-conference is you know? It takes a long time to go through it all and go through every single [component]… I had an instance where, if I'd said, 'Okay, do you have a[n American] flag in the room?' They missed that. They didn't see it, but I had American flag in the room you know, I mean it's not…and you go through every single thing.

I think the dialogue you have at post observation is important because I know in my first [observation] something was suggested and we had done it like a couple weeks ago. It pulls some things together or we were waiting for another time to do it at the end of the week because maybe it was half day. So [the observers] accept that. It's good to hear the

> suggestion when you know you either have done it, or you are getting ready to do it, they just didn't see it there.

The use of data to defend one's scores or argue for a change was apparent in Year 2 focus groups, as well. Many teachers across districts reported instances that they had used data provided as a decision point for ensuring that they received a score that they believed was accurate to their performance. Additionally, they continued to use the feedback provided for them to highlight areas in which they believed they had been under-rated due to observers either misunderstanding or missing teacher behaviors.

The tendency to use data in this way is not unexpected. The pilot program included higher stakes including the potential loss of tenure and employment for poor evaluations. As such, teachers would be wary of scores that they felt were not indicative of their true skill or performance. Nevertheless, this type of data use was more commonly mentioned by teachers in interviews than using data to improve teaching.

Seeking to Improve Future Observation Ratings

The second way that data was perceived to be used as an understanding of the observation protocol was that some teachers used the data to determine ways in which to get a higher score on future evaluations. This may be called "gaming the system." In one district's Year 2 focus group, secondary teachers described ways in which they and their colleagues could potentially ensure a "distinguished" rating on questioning techniques without actually affecting quality of teaching. In fact, the teachers speculated that this type of manipulation of their performance may actively harm students' learning, and as such were not intending to attempt it. This seems to be the extreme side; data use of this

type was not always as Machiavellian in its implementation. Instead, teachers were often more concerned with scoring well on the evaluation as a reflection of their own value. A key survey question illuminating this type of data use was asking teachers in Year 2 whether they had changed their teaching to get higher ratings. Table 4 shows the results of this question.

Table 4

Changing practice to gather higher observation scores, Year 2 (n=2593)

	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
Since the introduction of the new observation system, I have changed my instructional methods to get higher ratings	4.90%	36.91%	30.74%	19.44%	8.20%

The results here are, again, similar to the number of teachers reporting that they had improved the classroom environment and those who reported increased quality of teaching. Once more, however, teachers in the focus groups brought up this data usage with much greater frequency than improved quality of instruction.

The roots of this issue seem to come from the implementation of the observation protocol. Of particular concern to teachers was their understanding of how the top evaluation scores, most often called "distinguished" and scored as a "4," were going to be

applied. Focus group participants repeatedly spoke of being told that receiving a "distinguished" rating was, as several teachers recalled, "Like Hawai'i. You might visit there, but you won't live in it." Teachers expressed enormous vexation over being told they would likely not be able to achieve the highest possible ratings on their observation. Only seldom were the actual criteria for the achievement of "distinguished" discussed. Even though the rubric description of a distinguished behavior in a given category is extremely challenging to meet, teachers did not see the presentation as explaining that. Instead, they felt they were simply not going to be given the top score regardless of their performance.

For many teachers, this bone of contention was a starting point for using the observational data as an opportunity to learn how to get better scores. In six of the Year 2 districts teachers noted that the usefulness of feedback was linked to better understanding what administrators were looking for in assigning ratings. In the same way that teachers were concerned that they would not be able to be given the highest possible scores, teachers expressed interest in understanding just how raters were going to score observations. As one teacher put it, "The classroom observation tool, I think it's very lacking saying as far as being at the extreme end. I don't think that it gives you… it doesn't delineate exactly what the teachers need to do or like some exemplars or it's not explicit enough." This is a typical expression of teacher interest in understanding the rubric, and it does not focus on improving instruction but instead simply scoring well. Another teacher had experienced the post-observation conference as mainly useful in learning to score better. "My first observation came as a recollection but my post conference, the entire 45 minutes, 40 of it was spent on the logistics of the McREL

system and maybe five were spent on actually my lesson and me as a teacher," she recalled. While this teacher showed some regret that there was too little time for unpacking other aspects of the lesson, she was using that time to process and use information about the observation to understand the protocol better. In all, teachers across districts took the opportunity to learn about the new observation protocol.

Some teachers were more direct in their approach to this type of data use. As one such teacher explained,

> My standard hallway joke now is, for example, I signed up to be on this committee and the principal will say thank you. I said, 'It's okay,' he checks off another [protocol] box because, it goes. It's another artifact. Just the fact that I sent an email saying, 'I'll do this,' now I check off another box. That's gaming the system, that doesn't really do anything but it lends itself to that after a while.

This teacher's jest about how the system was working in his district betrays a more direct use of the data to make sure that scores are high. He jokes about using inconsequential interactions to bolster his scores, but the mindset of protecting scores by noting these small behaviors for future use is displayed.

The metaphor of "gaming the system" was also directly used in a different district in a Year 2 focus group. Here, a secondary teacher reflected on his feeling that the data provided and the protocol lent themselves specifically to "gaming the system." In this district, specific emphasis was being placed on evaluating teachers' use of questioning techniques that encouraged higher-order thinking. The teacher stated that it would be possible to ask a large number of "higher-order questions" in a lesson that was being

observed without affecting student learning. A colleague in the focus group agreed with his assessment. The teacher explained that he had not tried this approach, but that he and his colleagues had discussed it frequently. In these cases, teachers expressed a specific interest in using their knowledge of the protocol and previous scores to garner better scores in the future.

Another teacher, a colleague of the teacher who first used the "gaming the system" metaphor, responded to his comment, saying, "It's putting us, it's putting teachers, I think, what you said, telling us we are getting to the point where it's what do you want us to do we'll do it…." Here, there was less a sense of "gaming the system" and more a sense of being required to adopt new behaviors. Yet, she still displays willingness to follow these new requirements in order to get along in the new evaluation program. This is a subtle difference, but one that can be noted. Not all teachers seeking higher scores in the future are actively trying to –or at least contemplating ways to –manipulate scores. Some are simply complying with new regulations within their district bureaucracy. Though a reasonable action for teachers to take, this is another way in which teachers are using data to improve scores on future evaluations.

While it is clear that teachers more frequently saw themselves as using data for purposes of understanding the protocol, protecting their scores, and getting higher scores in the future rather than changing their teaching to improve pedagogical quality, at least one argument could be made that this is not a negative outcome for the program. Since the protocols used in the pilot contain teaching behaviors shown by research to be effective at promoting student achievement, even teachers who adopt more of these behaviors for the sole purpose of scoring better on the evaluation may find themselves

teaching more effectively without aiming for it. If one agrees that the behaviors in the protocols constitute good teaching, even if teachers are "gaming the system," the system may, instead, be equally gaming them, as well. Naturally, there are a number of factors that complicate this argument, such as frequency of implementation and overall quality of the protocol behaviors.

Data Non-Use

As shown, teachers in the pilot program reported using data in at least two different ways. More often than either of these data uses, however, was the failure to use data. As shown in Table 5, over half (54.75%) of respondents in the Year 2 survey chose either "neither agree nor disagree," "disagree," or "strongly disagree" when asked about whether they had improved the learning environment of the classroom and simultaneously chose one of the same responses when asked about changing their instructional quality. This means that more teachers report not changing their patterns of teaching due to data use than those that do in terms of what can be considered the desired ways. Moreover, of those teachers who did not claim to have altered their instruction or environment, 75.55% additionally answered that they had not changed their practice to get better scores. Tables 5 and 6 show these results. This means that 41.36% of teachers reported no change to their practice on these measures due to the feedback provided from observation. Once again, this number is in line with the number of teachers who report a change on any of the three measures posed in the survey.

Table 5

Cross tabulation of teacher responses to alterations in learning environment and instructional quality based on observational feedback. Year 2 (n=2608)

		Feedback has helped improve the quality of instruction				
		Strongly agree	Agree	Neither Agree Nor Disagree	Disagree	Strongly Disagree
Feedback has helped improve the learning environment	Strongly Agree	3.49%	0.31%	0.08%	0.00%	0.00%
	Agree	0.42%	31.40%	3.64%	0.69%	0.12%
	Neither Agree nor Disagree	0.08%	4.10%	26.96%	2.03%	0.19%
	Disagree	0.08%	0.73%	1.50%	14.42%	0.58%
	Strongly Disagree	0.00%	0.12%	0.31%	0.50%	8.28%

Table 6

Of teachers who reported changing neither their instructional practice nor classroom environment, whether they had changed instructional methods to gather higher ratings, Year 2 (n=1432).

Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
1.82%	22.63%	36.45%	25.63%	13.48%

In itself, 40% of teachers reporting essentially no use of the observational data is troubling, and suggests that there may be some barriers affecting their desire and ability

to use said data. These factors are the subject of the following chapter. As with the other forms of data use, qualitative analysis illuminates teacher feelings on the issue. The sentiment that the data given from observations was not useful was common among teachers interviewed in the focus groups, though mixed. While some teachers in most focus groups had experienced what they felt was useful feedback, many teachers did not find the data given them particularly helpful, nor was the new system better than previous attempts. This is emphasized in several comments from Year 1 teachers:

> No one has ever said to me –I continue to hear what a great job I'm doing but –'Hey, let's try this a little bit different or this might help you or there's some professional development opportunity for you.'
>
> Again, I'm not sure how that's supposed to help us grow
>
> I don't really find this to be largely different than anything else that we have ever done before.
>
> Leader: So how is this system as you're going through it now, how has it been useful to you? How has it helped you in your practice? The observation you've gotten, the feedback you've gotten, anything else about it, learning the system itself.
>
> [silence]
>
> Female speaker: Sounds bad doesn't it?

> I mean yes there was more technology and more evidence perhaps, but I think the conversations and what I as a teacher, as I'm growing in my field and becoming better at my craft, I feel like it wasn't so much different from this year to last year.

Outside comments like these, teachers frequently were quiet when asked in what ways they were finding the evaluation data useful. In Year 2 focus groups, this came through again. Though in one district teachers were particularly positive about the usefulness of feedback received, this seemed to be linked chiefly to the arrival of a new, respected administrator. This was the exception to the general trend of finding only a few teachers in a focus group who could report usefulness of the data. In two districts, teachers were unable to state any instances of data use from the observations, even when pushed to do so by the interviewer. In a third district, teachers received only the overall scores from their observation via email without verbal or in-depth feedback; they did not find this amount of data usable and, as such, did not use data. In such situations, some teachers did remark that the protocols had at least provided them an outlet for self-reflection, but this small detail does not change the overall lack of development reported by teachers in relationship to observational feedback.

This highlights a potentially disturbing outcome of implementation of the pilot observation protocols. Though some teachers reported data use, on deeper inspection is appears that large swaths of teachers found the data of little to no use to their development. Though it may be a reasonable action for teachers who believe the data to have limited value to essentially ignore it, is does signal a potential problem with the

evaluation system. This type of data non-use represents the failure of teacher evaluation to promote individualized formative growth.

Summary

Teachers had mixed feelings toward the overall usefulness of the evaluation data; these mixed reactions led to a variety of data-use actions for teachers. The types of data use reported fell into three main categories: improving practice, learning about the evaluation system and increasing future scores, and non-use. For each of these types, around 40% of survey respondents reported using data in this way.

Teachers using observational data to improve teaching is the goal of formative teacher evaluation, and, though self-reported, claiming this type of data-use is a marker of some level of system success. More objective measures of teaching quality need to be used in order to determine the extent of improvement, but this suggests that teachers are using data to change their practice in ways they believe to lead to stronger teaching. Qualitative analysis showed that this form of data use, however, was less used than the other types found in this study.

More common than purely using data for perceived improvement was using observational data to learn about the new system. This is an expected data use action, if not the intended one. As teachers were thrust into new protocols, it is natural for them to seek a deeper understanding of the system. Other teachers used the data as a tool to question the ratings they were receiving. In this way, teachers could argue for higher scores or defend the level of evaluation that they believed appropriate for themselves. Finally, teachers sometimes used data to, as one teacher put it, “game the system.” By taking the data they were receiving and applying their knowledge of the protocol used to

evaluate them, they were able to make targeted changes to attempt to ensure high scores in various categories on future evaluations. This finding may be less positive than the formative data-use intended by the system, but teachers who take this tack may ultimately improve their teaching if the behaviors they adopt to succeed on the observation protocol are indeed strong teaching behaviors.

A final data-use action was not using the data at all. Some teachers responded on the survey that they had not altered their teaching based on data received from observation. This finding is the most alarming, as it suggests that the work put into the formative side of teacher evaluation may be for naught. Looking into teachers' reasoning why they were not using the data, it appeared that the most common reason was that teachers simply did not find the received data useful. In this case, not using the data is a reasonable action, as there would be nothing there for teachers to use.

Each of these actions, especially non-use, suggest the need to look into what factors influence teacher data use in observation. The following chapter explores the ways some affective factors outlined in the conceptual framework, as perceived by teachers, play a role in their ability and willingness to use observational data.

CHAPTER 5: FINDINGS 2

Data Leak Reversal: How Teachers Perceive Data Use Factors

Whether teachers in the pilot program felt that their use of observation feedback was primarily for improving teaching, learning about the new system, or whether they did not feel they used the data at all, these perceptions were filtered through a number of different lenses. This chapter will examine teacher perceptions of several of these factors. By understanding the ways that teachers see these factors as helping or hindering their ability to use data, several hypotheses may be generated to develop a more substantial model of teacher data use in evaluation.

The research conducted for this work collected data on two main branches of the conceptual framework, system functionality and data quality. System functionality factors studied include observer capacity, time, logistics/scheduling issues, and trust. Data quality factors studied were observation protocol, specificity of data, validity, and reliability. Though the interplay of these and other factors creates a complex system for teacher data use, understanding how teachers view observational data through these filters may help to increase understanding of the process.

System Functionality

Teacher perceptions around system functionality were myriad. Systemic processes and procedures were a common focus within interviews and in the surveys, offering teachers an opportunity to vent frustrations and share successes. Four major factors affecting systemic functionality were commonly cited. In each case, teachers'

perceptions of the role of these factors help to shed light on their patterns of data [non]use.

Observer Capacity

Since the pilot program required schools and districts to increase the number of observations conducted for each teacher and additionally required the use of multiple observers, teachers were keenly aware of those who were observing them. Principals and other direct superiors continued in the pilot districts to be the main observers of teachers. This led to the first of two significant concerns of teachers around observer capacity: the content knowledge of the observer.

A second concern had to do with the pilot program requirement that teachers be evaluated by an "external" observer. The pilot program defined "external" loosely; any observer that was not based in the observed teacher's school was considered external. In most districts, this meant that administrators from the central office and principals from other schools in the district were used as external observers. Though these evaluators were closer to teachers than a complete outsider, teachers still harbored doubts about the contextual knowledge of the observer that affected their data usage.

These two themes, the content and contextual knowledge of the observer, arose as leading issues for teachers in terms of observer capacity.

Content Knowledge. Teachers were in agreement that being observed by someone who understood their area of expertise was an important part of a useful evaluation. Teachers in the Year 1 survey were asked a number of questions about the importance of content knowledge of the observer. Table 7 shows the results of these questions.

Table 7

Content knowledge questions in evaluation

	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree[3]
Evaluating good teaching in my subject area is different than other subject areas (n=2,293)	23.38%	34.98%	22.42%	15.66%	1.79%
A strong understanding of the pedagogy specific to my subject matter is essential for an accurate observation of my teaching (n=2,287)	29.12%	44.64%	17.14%	6.91%	0.57%
I would prefer to be evaluated by someone who understands my content area deeply (n=2,286)	24.43%	43.33%	27.11%	3.82%	0.61%
It is fair to be evaluated on my teaching by someone who is an expert on effective pedagogy even if they are not familiar with my subject area (n=2,280)	3.77%	27.81%	37.11%	19.21%	8.86%
Effective teaching is generally the same across all content areas (n=2,277)	7.29%	36.67%	25.21%	17.96%	10.14%

Teachers strongly believed that content knowledge of the observer was essential to the observation process, with 73.76% agreement with the question. Moreover, 67.76% of teachers espoused a preference for being evaluated by someone with content expertise. A minority –though, somewhat surprisingly, a significant one –believed that being evaluated by someone without specific content knowledge was fair and that teaching was

[3] Additional possible response was "Not Applicable"

generally the same across all content areas; 31.58% and 43.96% agreed with these statements, respectively.

In Year 2, 86% of teachers out of 2,702 respondents responded that it was either "Important" or "Very Important" that their observer had content knowledge in their area.

These results suggest that teachers see content knowledge of the observer as an important facet of an effective evaluation. When probed more deeply in focus group interviews, teacher opinions were even more decisive. "Yeah I would prefer someone more with understanding what I'm teaching because they could probably figure out why I'm doing it the way I'm doing it more than someone who knows me and who sees me like all the time," said one teacher. Others expressed similar desires. As one teacher said, "I think I would hand pick my observer to be a person who knows my subject area and how it should be taught because not all subject areas should be taught the same." Another put it this way,

> Yeah it's funny that you have to be, to be a teacher you have to be highly qualified in a certain area and to be an administrator for a certain area you don't have to have that background in that curriculum which is kind of a flaw in the system in general. It's not just this process but I feel like if you are going to be observing Math teachers you should have a good content knowledge up to the highest level of Math that they offer in that district or Science or whatever other field, or a language.

Teachers voiced these concerns not only across subject areas, but grade levels, also. An elementary teacher noted, "… this paper tells you that I'm certified in teaching K to fifth grade, so you know… And I mean, there's a huge difference between

kindergarten to fifth grade." The same was true of special education teachers, "I often feel like I need to remind administrators that I do have a first grade special ed class although I'm supposed to be, and I do, focus on the same curriculum," a special education teacher said. Throughout all the focus groups, teachers agreed that content knowledge, and pedagogical content knowledge, were desirable traits.

Two main reasons for these feelings arose in the focus groups, both related to data use. First, teachers believed that observers familiar with teachers' areas of instruction or grade level would be able to provide more useful feedback in the observation. One teacher hit this belief on the head:

> …but at our level I think there needs to be an element of content knowledge especially for the feedback and growth as a teacher so if you have someone who is not privy to the curriculum, that standards, the departmental objectives –there's national standards –I'm not going to receive the feedback that a supervisor of English could give me and I think that's greatly missing at the high school level. Because the administrators aren't certified in your content area, so to get feedback to make you a better teacher I think content knowledge is essential."

A teacher in another district shared similar thinking:

> I can tell you something I really appreciated as an educator was a couple years ago when we has a curriculum coordinator, which made me really happy because he was a science guy. But he would come to me with ideas and different things and it was more of that I felt like I was being taught and I was growing as an educator, whereas [the new observation system]

> does not give me that sense. This gives me a sense of stifling and not growth.

As did another teacher, saying, "…how are you going to get suggestions for improvement on how to teach Spanish when you're from outside the building and you don't even know Spanish…?"

Teachers clearly believe that observers with content knowledge will be able to give more useful feedback. The converse to this statement is important, as well. Teachers believe that observers without content knowledge will provide less useful feedback.

A second reason for teachers' preference for observational content knowledge was related to the accuracy of the observation scores. Teachers believed that their observation scores were potentially less accurate and valid when scored by someone without content knowledge; conversely, content specialists would reflect better on their evaluations:

> I think it's just the accuracy overall of the different job descriptions that we all have. I mean, everybody can't know every single thing, but I think it's very tough if you've had an administrator who, say, taught second grade, who has never taught at middle school or high school, and now they are coming in to observe you as a middle or high school teacher. … I think it's very hard to be accurate.

> We used to have a superintendent and she was an assistant superintendent even though that wasn't her title, and one was more math- and science-based and the other was language arts. [Teachers] never did as well when she observed

somebody who taught math or science, because she didn't understand what she was looking at, so she gave them all great. And when the man with the stronger math background went into language arts and he said they were all great because he didn't really understand what they were talking about.

For science it's different than music is different than kindergarten is different from first grade, and I think the accuracy gets a little grey because it's not the same.

For teachers, concerns about observer content- and pedagogical content knowledge played into both a formative usage, getting useful feedback to improve practice, as well as one of a more summative nature. Therefore data use, in the minds of the teachers, is affected by the level of content knowledge of the observer.

Contextual Familiarity. In addition to content knowledge, teachers were mostly united in their belief that an observer who understands the context of their classroom. Feelings about this issue typically crystalized around the use of external evaluators for observation. Table 8 shows survey responses from both years of the pilot program around the use of observers from outside teachers' schools. As a point of reference, 62% of teachers (n=2,915) had experienced at least one observation conducted by an external observer at the time of the Year 2 survey.

As with content knowledge of the observer, teachers see contextual knowledge of the classroom and a professional relationship with their observer as very important. This poses a bit of a conundrum for observation, as administrators from teachers' schools, presumably those with the most contextual knowledge, will not likely possess the specific

Table 8

Teacher perceptions of familiarity with observer

Year 1

	A superior with whom I have a developed professional relationship	A superior with whom I am not very familiar	A teacher with whom I have a developed professional relationship	A teacher with whom I am not very familiar	Someone whom I have never met
I would prefer to be evaluated by (n=2,298):	79%	5%	12%	1%	4%

	Strongly agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
I would rather be observed by an impartial observer than someone who knows me for purposes of accountability (n=2,286)	3.37%	12.25%	36.31%	35.65%	12.42%
I would rather be observed by an impartial observer than someone who knows me for purposes of professional development (n=2,288)	3.37%	12.41%	36.31%	33.35%	11.76%

Table 8 (cont'd)

Year 2

	Very Important	Important	Neither Important nor Unimportant	Unimportant	Not at all important
How important is it to you that the person observing you knows your classroom well (n=2,712)	47.75%	32.15%	14.71%	3.24%	1.40%

pedagogical knowledge for diverse areas of teacher expertise in their schools. Nevertheless, teachers expressed strong desire for both these facets to be present in observers. Once again, focus groups provided more direct insight into this particular factor.

The issue of external observation was an emotional one. As one teacher said, "And as far as having strangers come in the room, I think that it *just completely destroys the whole observation process*… You know at least [known administrators] have familiarity with your teaching style or your students…." [emphasis added] This teacher's vehement rejection of the external observation is the key concept in framing teachers' views about being observed by an outsider. Another teacher had a similar comment: "I don't see how that has any fairness as to what you do for these kids on a daily basis; you have some outsider come in and evaluate you… that's impossible." There was a distinct air of negativity throughout the discussion of using an observer with no relationship to the teacher or classroom. While not all teachers strongly opposed external observation, teachers in all districts agreed that in most cases, an observer with whom the teacher had a working relationship was strongly preferred on many levels. One teacher considered

using an evaluator who had a working relationship to be given of a strong evaluation system, saying,"…If you're going to take it seriously, the first thing I would have is the administrator who knew you well be part of the process. It makes sense."

Overwhelmingly, the concern about external observation boils down to concern about contextual understanding. Teachers do not believe that an external observer unfamiliar with the teacher's style or classroom dynamic can make a qualified observation. Several comments illustrate this concern.

"I also have another concern, who is giving you this review? If it's your principal that is fine; if it's someone who doesn't know you there's a concern because he doesn't know your teaching methods, your strategy. He doesn't even know the type of children you teach."

"Well, I think whoever is observing us more importantly has to know our student body."

"I know he was basing a lot on things that I had done and there were questions in there that you can't see. Do they realize that? Do they understand that?"

Teachers repeatedly spoke about the differences in individual students within their classes, and voiced concerns that a stranger to the classroom may fail to understand the intricacies of interpersonal relationships and discipline within a single lesson. Teachers gave many examples of situations in which a principal, subject-area supervisor, or other direct superior would understand the context of student dynamics more clearly than an outsider.

Contextual understanding of the classroom, much like content knowledge of the observer, was felt by teachers to be requisite for a useful evaluation. These two elements

of observer capacity must, then, be included in exploring how teacher data use is affected by system functionality factors.

Time

Another systemic factor cited by teachers as affecting their ability to use observational data was the time required for the work. Increasing the required numbers of observations, as well as adopting a new, more rigorous observation protocol added to the tasks teachers were asked to accomplish during the school year. Teachers in Year 1 believed that the new evaluation took more time than was necessary, as shown in Table 9.

Table 9

Whether the observation system takes too much time, Year 1 (n=2,319)

Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
25.66%	38.64%	23.98%	10.61%	1.12%

Even more striking than the 64.30% of teachers who agreed that the new system took too much time was that only 11.73% of teachers disagreed with the statement. This shows that most teachers felt time pressure placed on them by the new system. One way to explain this finding is due to the newly implemented system itself. Since the new observation protocols were more thorough than the systems they replaced, teachers would need to do a good deal of learning in order to understand the new expectations placed on them. This was reflected in the number of teachers who noted that they had changed their teaching in order to get higher observation scores or used data to learn more about the system particulars. In addition, teachers were asked to complete pre-observation

paperwork and sit for conferences with raters both before and after the actual observation, all of which added to the time teachers devoted to the new observation system.

Teachers in Year 1 focus groups mentioned the time that the new system was taking frequently. “I can’t save a lesson just you know just that I need to right up the whole thing again and spend, ‘cause you know you want it to be perfect, you want to see whether just teach what they’re looking for a distinguished teacher, I mean it took a lot of time. But I guess it didn’t mean anything,” said one teacher, reflecting on the time needed to schedule and develop a strong lesson for evaluation.

A teacher in a different district observed, “I know so many teachers who have spent countless hours on this rubric which I don’t think you should be as a teacher you should be concentrating on your content and how to improve your lesson the next day.” This teacher felt that the time taken for the evaluation might be better used in other ways more directly related, in her view, to the classroom. This suggests that data use in this case might be more limited as the teacher diverts attention to other information points for teaching improvement.

Even when teachers voiced opinions positively around use of the data, time was a noted concern. As one teacher noted,

> What I found was, as in, I mean I like the pre planning and the post-planning pieces because you can really give a picture and you can go back and reflect. But quite frankly as a teacher with so much stuff that we need to do, I found it a little tedious. And I was like I could really sit down and take twenty minutes to now do the pre and then a lot of questions seem

> repetitive. And so on one side it's good because you know it's kind of circling back to where you started with it but on the flip side, I did find it a little tedious. Like I was like come on like I don't have time for this you know.

In this case, even an opportunity to use self-generated data was found to be too time consuming, which may limit teachers' data use. For Year 1 teachers especially, time appears to be a significant issue.

In Year 2 focus groups, time in this fashion was less often mentioned by teachers. While teachers in most districts acknowledged that time was a factor in their use of the evaluation system, more of the comments were focused on time as it related to logistical and scheduling issues rather than directly to teachers' ability to use data. This was consistent across both districts that had participated in Year 1 of the pilot program and those new to the pilot in Year 2.

The relative cooling of time as a data use factor in Year 2 may be explained in three ways. First, for districts that were in their second year of implementation, there was more comfort with the new system, likely making teachers feeling less stress about the work needed to prepare for and receive data from observations. Secondly, Year 2 learning in all districts was more highly focused around the use of student growth data as an evaluation measure, something that had been mostly neglected during Year 1. This focus on a different facet of the evaluation may have caused observation time pressures to be less evident or problematic to teachers. Finally, some districts new to the pilot in Year 2 were more economically advantaged than their peers from the first cohort. This was not

universally true of the second cohort of districts, but the relative capacity of these higher-income districts may have affected teachers' feelings about time effects in data use.

Logistics/Schedule

The logistics of scheduling and completing the various phases of the new observation protocol also arose among teachers as an issue that affected teachers' ability to use data. In this case, the specific reasons mentioned by teachers were tied strongly to teachers' ability to even receive data for potential use: delays or cancellations of scheduled observations, and delays in, or failure to receive post-observational data in a meaningful way. As noted earlier, teachers reported in the Year 2 survey that post-observation conferences were one of the most frequently received sources of observation feedback. Specifically, slightly more than half of the 2,711 respondents to this item, 51.27%, claimed to have received this type of feedback "Very Frequently" or "Frequently." Of course, this means that the remaining 48.73% responded that they had received feedback less than frequently, with 16.34% claiming they "Rarely" or "Never" received feedback from post-observation conferences. Table 10 shows these results.

Table 10

Frequency of feedback from post-observation conferences, Year 2 (n=2,711)

Very Frequently	Frequently	Sometimes	Rarely	Never
12.54%	38.73%	32.39%	10.92%	5.42%

This was a noted problem throughout districts in both Year 1 and Year 2 of the pilot program. In one district in Year 2, observers had ceased conducting post-

observation conferences altogether as a way to try to complete more of the required observations; feedback was only delivered via electronic score reporting. This would significantly hamper teachers' ability to use observational data since they would have limited access to such data with which to begin. In other districts, teachers reported extended lag between the actual observation of the lesson and the post-observation conference. Though some teachers reported having their post-observation conference with their evaluator with the pilot-specified 5 days from the lesson, teachers often noted receiving feedback anywhere from 3 to 8 weeks after their observed lesson. Some, as in the district cited above, received no post-observation conference at all. This kind of timing issue was common among Year 2 focus groups; the frequency of reported occurrence varied by districts. In Year 2 focus groups, lagging feedback was the most common time-related issue cited by teachers, rather than the overall time taken to work on the observation. In at least one district in Year 2 the gap between observation and feedback was noted to make the feedback "meaningless."

Teachers experiencing this sort of issue found this to hamper data-use; their perceptions and recollections of the observed lesson were diminished by time lapse. This is highlighted by an exchange in one of the Year 1 focus groups:

> Female Speaker: What I found difficult was the reflection after in my building when they would post the observation, we would talk about it. It would often be a couple of weeks after the actual observation occurred so by the time I was required to write the reflection, I didn't even remember specific things that happened in the classroom. So that was pretty difficult.

Leader: So it's been a timeline between your observation and the…?

Female Speaker: Yes.

Female Speaker: I was going to say the same thing but I thought the process was a lot longer. I mean I got observed one month and then I didn't get my observation the same as that thing. I would completely draw blank cause I forgot to be honest with you. And I'm thinking observation is just a little snippet of our lesson so it's kind of like it's hard to reflect on something when it was so long ago?

Teachers felt that having too long between observation and feedback was a barrier to effective data-use as they could not accurately recall their lesson and thus struggled to be able to implement feedback for improving their teaching.

A second logistical issue was teachers' frustration with delays in the observation schedule. Teachers felt a great deal of frustration with preparing for an observation and subsequently having it delayed or cancelled. Two teacher comments from different districts showcase this frustration:

"Well on the [data system] I spent two and a half hours 'cause I wanted to make sure everything was correct and no one ever read it. So I spent all this time on it and then nobody came the day I was supposed to have my observation."

"… You put all this prep work into getting a lesson prepared, and you do the pre-conference and then the administrator has to cancel, like, that really stinks."

Teachers found that putting off a prepared observation meant that they would not score as well on the observation when it actually happened, and therefore felt more defensive about their session. While there were no comments directly addressing the data-use effects of such feelings, it may be reasonable to presume that negative emotional inputs may adversely affect teachers' ability to use data effectively.

Trust

One emotional issue that was addressed was that of trust between teachers and evaluators. As noted above, teachers did not show a great deal of trust in the opinions of evaluators with whom they were not familiar. At the same time, there were a significant number of teachers who voiced their preference for a superior with whom they had a professional relationship to evaluate them. As shown in Table 11, the Year 2 survey reflected a similar sentiment among teachers, who evinced trust in both their direct supervisors –principals –and in one another. In each of the questions exploring trust issues in the survey, no fewer than 60% of respondents agreed that they trusted their principal and colleagues.

Teacher trust in those around them may suggest that teacher would feel more comfortable in receiving observational data –particularly if that data was challenging to teachers' perceptions of their practice –and therefore have an easier time in processing

Table 11

Agreement levels with statements regarding trust, Year 2

	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
Teachers in this school trust each other (n=2,635)	13.51%	47.13%	22.69%	13.13%	3.53%
I trust the principal at his or her word (n=2,646)	28.65%	37.04%	18.86%	7.86%	7.60%
It's OK in this school to discuss feelings, worries, and frustrations with the principal (n=2,644)	25.91%	36.95%	18.08%	10.85%	8.21%

and using evaluation data. The idea of trust certainly applies in deeper consideration of the use of external evaluators.

An additional issue of trust arose in understanding teachers' perceptions of the purpose of the pilot program. Strikingly, teachers simultaneously held both formative and summative purposes as a part of the evaluation pilot. This means that teachers experienced a constant struggle to interpret the meanings of their scores and were fearful of consequences that may result from negative feedback, while generally appreciating the feedback received in the new system.

Teachers were clear that they believed the primary purpose of the new evaluation was to help to develop them and improve their teaching. Teachers mentioned frequently that the new system was to provide more comprehensive and developmental feedback to their practice. "I think the good thing is that it does promote discussion, it does promote dialogue; it promotes, also, teacher growth…" said one teacher. In a different group, another teacher said, "I like the pre-planning and the post-planning pieces because you can really give a picture and you can go back and reflect." These comments demonstrate

both an attitude of professional development and the feeling that feedback was provided to help teachers improve their practice. These feelings held true in other focus groups, as well:

> … It promotes teacher growth because you get to be able to really see your goal setting, where you would like to be.

> Isn't the whole idea supposed to be to make us better teachers and to make our students more successful?

> I think the post-ob[-servation conference] is helpful, too, because it gives you a chance to reflect on your lesson and to…. Because we normally don't have time moving from subject to subject to sit back and say, 'Where did I go wrong?' or, 'What changed in the lesson?' or the positive things, the outcomes from the lesson. [This] gives you a chance to reflect on it.

> I think this process puts more on the administrators to make sure that we get better at what we do. I don't think that happened before.

> I think no matter what observation you put into place it's all basically with the same end result. It is supposed to eventually lead to areas of professional development.

These comments illustrate that teachers feel that the protocols now in place in their schools are meant to help them develop as practitioners.

Teachers' beliefs that the goal of evaluations was primarily formative contrasts with their suspicions of the summative aspects of the program. While teachers speak of the feedback and reflective qualities to the observation protocol, they simultaneously harbor deep misgivings about the use of the same evaluation for summative –and what are almost always perceived to be punitive –purposes. While these statements were less prevalent in the focus groups, this dual belief in the benevolence and malevolence of the system suggests that trust in the system is not concrete; this likely affects teacher data use.

One teacher explained, "I almost feel like this system is almost set us up to be brought down a notch. I think as if somebody up above is like, 'We got to bring these people down a notch. They almost think they are too good,' and this is the system to do it with. Just a standby like we are not as good as we think we are because the system has set us to fail." This statement encompasses a great number of fears, and illustrates a deep lack of trust in the purposes of the district and state. Another teacher echoed this statement: "…if they are going to review it or if they are serious about it being a useful too, then that would be good. If it's a tool to find out that we are all deficient then obviously that's a waste of everybody's time." A third teacher embodied the simultaneously formative and summative question, saying, "You know, what is the point of it exactly? I mean, if it's to improve education, that's one thing; if it's to cut us all out, that's something totally different. And ultimately, who is looking out for the best interests of the students?" This teacher had yet to determine what she perceived as the overall purpose of the pilot program.

Ironically, concern about the summative aspects of the system did not extend to teachers' overall job security. Teachers were asked whether they felt that the evaluation system would affect their chances of receiving tenure –for non-tenured teachers –or losing it –for tenured teachers –in the future. Table 12a, below, shows that tenured teachers did not believe, in general, that their job security was greatly affected by the new system, despite their belief that the system was partially summative/punitive in nature. Table 12b suggests the same is true of non-tenured teachers. In this case, the formative belief seems to trump the summative side.

Lack of trust and fear of reprisal from negative summative reviews –be they tangible effects like loss of tenure or motivational effects from receiving negative feedback –may help to explain teachers' resistance to external observation, and is a factor to be explored in its own right. If teachers fear contextual misunderstanding will lead to poor reviews, then their mistrust may extend to data use; a teacher who does not trust the system may in fact use data less than those convinced that the purposes of the data are for their growth. Moreover, lack of trust may encourage teachers to attempt to exploit loopholes in the protocols in order to game the system and receive scores that make them feel more comfortable, rather than critically reflecting on their practice and endeavoring to create change.

Table 12a

Tenured teachers' response to potential summative consequences of observation, Year 2 (n=2,268)

	Very Likely	Likely	Neither Likely nor Unlikely	Unlikely	Very Unlikely
Do you think you have a chance of losing tenure under the new system?	3%	6%	23%	31%	37%

Table 12b

Non-tenured teachers' response to potential summative consequences of observation, Year 2 (n=440)

	Greatly improve my chances	Improve my chances	Neither improve nor reduce my chances	Reduce my chances	Greatly reduce my chances[4]
How do you think the new evaluation system will affect your chances of getting tenure?	4%	22%	50%	4%	3%

System Functionality Summary

It is clear that system functionality factors play a significant role in teachers' perceptions about how observational data can be used. Issues of observer capacity were foremost in the minds of teachers, who saw the lack of these things as directly negative in terms of generation and use of meaningful observational data. On the other hand, teachers expressed trust in their principals, suggesting that data use may be more likely in

[4] "Don't know" was an additional response

situations where the principal is involved in the feedback process. Finally, time was a factor in data use in that teachers, when feeling harried by the demands of the program, were less likely to use data fully, and in some cases were unable to even interact with useful data when it wasn't delivered. This led into frustrations with scheduling and logistical issues such as lag time between observation and post-observation conference and missing scheduled observation sessions. Each of these factors were noted by teachers as affecting their use of the evaluation data, and tied into one another in a number of ways. Overall, it seems likely that system functionality significantly affects teacher data use and the data use process.

Data Quality

Data quality was the second branch of affective factors for which data was available for this study. In general, teachers felt that when the observational data was perceived as of higher quality, data was more meaningful and more willingness to use said data was apparent. The following section will outline the way that teachers saw specific data use factors from the conceptual framework as relating to their data use. The observation protocol, specificity of data, validity, and reliability each were noted as having an effect in this area.

Observation Protocol

In each of the pilot districts, teachers experienced the implementation of a new observation protocol. Some districts selected a new protocol based on prior experience, e.g. selecting the Danielson model because the district already used a form of Danielson. In these cases, the districts adopted updated or more complete versions of the protocol,

Table 13

Agreement levels with questions pertaining to capabilities of the observation protocol

	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree[5]
Year 1 (n=2,319)					
The district's system for assessing teachers clearly separates accomplished from unaccomplished teachers.	4.18%	23.33%	30.19%	21.69%	11.47%
Year 2 (n=2,739)					
The new teacher evaluation protocol can be used to clearly separate accomplished from unaccomplished teachers.	2.12%	27.35%	26.47%	27.20%	12.89%
The criteria in the new teacher evaluation protocol can be used to appropriately assess teachers' performance.	1.42%	28.95%	27.93%	24.90%	13.03%
The new teacher evaluation protocol can be used to assess what is important about teaching.	2.52%	28.48%	27.78%	23.04%	14.53%

resulting in new experiences with more familiar protocols. These experiences also encompassed new technology being used for observation, such as the use of tablets to make notes and ratings in real time during observation sessions. Survey data suggests that teachers were generally cool on some of the touted characteristics of the new protocols, as shown in Table 13. In both years of the pilot, teachers were asked whether the new protocols were capable of separating strong from weaker teachers. More teachers

[5] In Year 1 survey, additional possible responses were "Does not apply" and "Don't know." In Year 2, "Don't know/Not sure" was an additional response

disagreed with this statement than agreed in both years, with more teachers expressing disagreement in Year 2 than in Year 1. Additionally, more teachers disagreed in Year 2 with statements around whether the protocol could appropriately measure performance or that the protocols were focused on the important aspects of teaching.

In focus groups, the idea that the protocols did not always focus on the most important aspects of teaching arose again. Teachers remarked in several districts about the importance, and implied usefulness, of the large variety of data collection points for observation. Several comments from teachers suggested that they were not certain that all the data points were meaningful. Reasons cited ranged from the feeling that there were too many items to observe to the feeling that certain rated behaviors were not central to the goal of strong teaching.

Said one teacher, "At times it's a little over the top. I worked at a different district where a superintendent once said, 'I'd rather have fifteen rules that we implement than a hundred rules and we only implement twenty of them.' So, a lot of it looks good but is it all practical? I'm not always positive of that." Another teacher spoke to her feeling that the protocol didn't hit the parts of her lesson that were important, saying, "I think that it is kind of cookie cutter; it has to fit into this package so perfectly. Every day in teaching there's something different and that might not apply every single day in teaching because of the nature of our subject areas." And in a third district, the issue of looking at a teacher's involvement outside the classroom, common to each of the protocols, was called into question. "I feel like that the [observation] system is too focused on if I coach, if I'm an advisor, how many times I see the kids outside of my classroom, instead of

they're with me, how are we going to engage them, how can we make this better," said one teacher.

This sentiment was not unanimous across districts. Some teachers noted factors that they felt were particularly pertinent to strong teaching. These were often the direct teacher interaction items, such as questioning techniques and student engagement. These were noted in Year 2 focus groups and in Year 1 as being important parts of teaching. This lends some balance to the feeling that the protocols did not measure the important parts of teaching. It is safe to say that teachers general felt that the new observation protocols hit some of what they felt was important in teaching while also creating "noise" by generating data that was not germane to their experience.

This was another case in which teachers acted to filter the feedback. This filtering is both a part of the process but also may suggest that the protocols may present a barrier to data use. If teachers do not agree that the items on which they are receiving feedback constitute the important aspects of teaching, there would be little potential benefit for them to use the data in any way, save to potentially gather higher scores on future evaluations. In any case, the protocol and teachers' perceptions around it and its capabilities do seem to play a role in affecting data use.

Specificity of Data

Only a small amount of data was collected regarding teachers' perceptions regarding specificity of data, as this was not directly addressed in either teacher survey. Focus groups, however, provided some small insights into how the new protocols were perceived in terms of their ability to provide specific data to teachers. These comments tended to be centered on the protocols' use of four rating levels for each behavior.

Perceptions were mixed around these scoring levels. Some teachers responded positively, commenting that they liked that they were able to see the specific levels of each teaching behavior. This allowed teachers to see where their techniques and choices would fall on the rating scale. “One of the things I liked about it was… that the domains are very specific and pretty much anything falls into those categories,” said a teacher in Year 1, referencing the ability to look for the various moves she made in the classroom. A colleague agreed with her, and felt that the specific levels could be a source for potential data use, remarking, “… it’s good to see that you’re doing well or poorly, and then if you’re doing poorly in an area, you could see why or what you’re missing.” In a different district, another teacher appreciated that there was more data to be filtered by the protocol. “Because they’re assigning everything to a certain standard, so the more you write, the more you have, the more information you have to show that they can assign to those things,” she said. This teacher felt that having more evidence of success or challenge in each area of the rubric would provide a firmer basis for her observational data. In a third district, yet another teacher appreciated the increased specificity, saying, “And it’s more specific because we have like the different components and the different domains and we know what they’re looking for and everything is going to be supported by evidence… It’s not going to be as subjective as in the past.”

Though not a topic that was at the forefront of teachers’ minds, teachers did seem generally appreciate of increased specificity in the data provided by the new protocols. They seemed to feel that these specifics would benefit both their ability to use the feedback provided to them as well as create a greater accuracy for their overall evaluation.

Not all teachers had completely positive reactions to the increased specificity. Some teachers found the some levels of the rubric to be unreasonable or that there were too many behaviors to be scored. As one teacher put it, "When you are truly introducing something for the first time, it's hard to really hit all those areas and do all of those things that the rubric is asking you to do because it's completely new to the kids." Other teachers felt similarly, and were concerned that their scores would suffer if they failed to address each of the components of their particular evaluation system. A few others were concerned about not being able to meet the top levels in some categories due to student challenges. In regards to the Danielson item about higher-order questions, one teacher responded, "I have [students] that are four slash five or like just turned five in August and will turn six again. So they are five years old. You can't ask them a three level question if they have no prior knowledge to that."

In these cases, while teachers did not seem to appreciate the specificity of the protocol, their concerns focused on their ability to achieve the "distinguished" rating, or the analog in their district's protocol. These concerns were not easily related to their ability to use, and the process by which they used data. While it is possible that these concerns may play a role in teachers' feelings of trust in the system, which may in turn have an effect on data use, this is an ancillary effect that was not measured in this study. Overall, teachers in general felt that the greater specificity offered by the pilot program observational data provided made data use easier. Though relatively minor in terms of how frequently specificity of data appeared among teacher comments, this finding is significant since it represents an aspect of the program that teachers felt brought a positive effect rather than one with positive potential or negative effects.

Validity

Teachers were also somewhat quiet about the validity of the system directly. Teachers are not psychometricians; they are concerned less with whether the pilot program used technically valid measures than how the observation and data affected them as teachers. Therefore, teachers were not asked in the survey about their perceptions of the validity of the observational data. Teachers were, however, asked how they viewed the protocol's level of fairness and accuracy. These questions are used here as proxies for teacher feelings of validity. If teachers feel that the data they receive is both accurate and fair, it is safe to assume that they would feel that the data was valid. Table 14 shows the results from these questions in three ways: from the Year 1 survey, the Year 2 survey, and the Year 2 survey filtered to responses from only teachers from districts that had participated in both years of the pilot program.

In Year 1, more teachers felt that the observation system was accurate and fair than those who disagreed, though there was not a great disparity between the two sides. About one third of teachers were neutral in their perceptions. While not overwhelming, this suggests that a fair number of teachers in Year 1 had neutral to positive feelings about the validity of the protocol. This balance shifted in Year 2. Here, fewer teachers believed in the system's fairness, and more disagreed with the questions of fairness. This was also the case in Year 2 for cohort 1 teachers, meaning that teachers had become more negative on the system's fairness.

In terms of accuracy, roughly the same percentage –around 30% –of teachers agreed that the system generated accurate assessments in Year 1 and Year 2. Slightly

Table 14

Agreement levels with questions pertaining to observation accuracy

	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree[6]
Year 1 (n=2,319)					
The district's system for assessing teachers generates accurate assessments.	4.01%	26.30%	33.68%	19.53%	10.31%
The district's system for assessing teachers is fair.	4.44%	32.77%	31.69%	16.30%	9.96%
The new observation system provides a fair picture of my teaching	4.61%	31.22%	35.66%	19.53%	8.97%
Year 2 (n=2,739)					
The new teacher evaluation protocol can be used to generate accurate assessments.	1.83%	28.84%	29.65%	23.04%	11.50%
The teacher evaluation protocol for assessing teachers is fair.	1.39%	26.32%	29.46%	25.85%	13.33%
The teacher evaluation system provides a fair picture of my teaching.	1.50%	24.79%	29.86%	26.07%	14.13%
Year 2 – Cohort 1 Districts (n=1,574)					
The new teacher evaluation protocol can be used to generate accurate assessments.	2.10%	31.45%	31.58%	21.09%	9.59%
The teacher evaluation protocol for assessing teachers is fair.	1.46%	28.14%	32.08%	24.21%	10.99%
The teacher evaluation system provides a fair picture of my teaching.	1.59%	27.19%	31.51%	24.40%	12.39%

[6] In Year 1 survey, additional possible responses were "Does not apply" and "Don't know." In Year 2, "Don't know/Not sure" was an additional response

more teachers in Year 2 disagreed with the system accuracy, as fewer were neutral or undecided. For teachers in districts who had participated in both years of the pilot, a higher percentage agreed with the accuracy of the protocol in Year 2, while disagreement levels remained almost constant between years.

These findings suggest that teachers were balanced on their perceptions about validity of the protocol in these terms. Teachers with two years of experience in the program were slightly more likely to believe in the system's accuracy, perhaps signaling positive movement in their perceptions of the observation protocol. It is safe to say that teachers as a group held varying opinions whether they felt the observation system was valid.

Teachers in the focus groups were more consistent in their feelings about the fairness and accuracy. When asked whether they felt that the observational data painted an accurate picture of them as teachers, most felt that it did. In no case did teachers respond that they felt that their observations had produced wholly inaccurate or unfair results. This was the general consensus of teachers across both pilot years and districts.

These feelings, while generally consistent, were not unanimous. Teachers of non-core subjects such as music, special education, or ELL sometimes commented on concerns that the protocols were potentially less valid for their subjects. One teacher was particularly vocal that the protocol had "nothing to do with music." This teacher clearly felt that the protocol was not valid for him, and he made clear that he had no intention of using the evaluation data. This suggests that if teachers perceive the system as invalid, data use will suffer.

Yet this type of response was the exception, rather than the rule. Indeed, other teachers of specialty areas including music and foreign language did not express such extreme reactions to the program's validity. And while special education teachers often posed scenarios where they felt that their students and teaching may be misunderstood, and ELL teachers worried about how to meet the demands of the protocol with their limited-English students, these teachers tended to remain mostly positive about whether the observations were fair and accurate.

Thus teachers tended to view the validity of the system in shades of grey. Though posing questions and clearly not fully bought into the system, teachers who had been evaluated tended to find that the system produced reasonably accurate results. Teachers did not often reflect on how the validity of the system directly affected their data use except in the extreme negative cases, where feelings that the system was invalid had an extreme effect on the teacher's use of data. This at least suggests that the converse – increased perceptions of system validity –may have a positive effect on data use.

Reliability

Closely related to validity was the reliability of the observational data. Teachers were moving toward more negative attitudes about accuracy and validity in Year 2 of the pilot while remaining reasonably balanced; they were more uniformly aligned on the issue of reliability. Teachers expressed doubts about observation reliability. These doubts were tied to the issue of being rated by multiple observers. Survey responses from Year 1 suggest that teachers believe that the evaluators coming in to their classrooms have received enough training and that they possess the knowledge needed to perform the

evaluations, as shown in Table 15, below. About half of respondents agreed with these statements, while only 15%-18% disagreed.

Table 15

Agreement levels with questions pertaining to observer training, Year 1

	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
I feel that the evaluators in my school have the required knowledge and competencies to appraise teachers. (n=2,294)	9.59%	44.55%	27.55%	12.16%	6.15%
I feel that the evaluators in my school have received adequate training to perform their job. (n=2,293)	9.68%	39.69%	34.80%	10.60%	5.23%

Despite this, teachers felt that the ratings they received were mostly inconsistent, with many teachers commenting on ways in which scores differed from observer to observer. In one district, when asked for positives and negatives about the new observation program, one teacher responded,

> My colleague, for example, had been observed by one of our administrators. She got in two of the categories ‘outstanding’ and she got some feedback from the administrator. Three days later, she had another observation, incorporated the feedback to the exact same lesson and a second administrator observed her. It was kind of a drastic difference in the observation. And so two same exact lessons with feedback incorporated I would imagine only could get you a little bit of a better

> observation, but that one was lower. So that really makes me question the, how much objectivity, you know, kind of plays into, it's still also a factor of it, right?

This underscores how data use might be affected by real or perceived reliability issues. In this case, the teacher tried to use observational data to improve her teaching, only to find the second observer found her performance poorer than before. While a complete account of the overall evaluation is not present, the teacher relating the story had clearly internalized reliability issues as relating to data use. Here, lack of reliability created a confusion point for data use.

Another teacher in the same group noted, "…When they presented this to us they said, 'Oh, it's objective,' and they stressed that very much. I don't think it's objective." Both these comments demonstrate the overall feeling that the ratings they receive are not reliable from observation to observation or from observer to observer.

In a third district, subjectivity continued to be a common thread. "It's still very subjective and not all staff members are being treated the same way," was representative of the feelings on reliability. Focus group teachers throughout both Year one and Year 2 frequently talked about the perceived failings of different raters scoring teachers inconsistently.

This sentiment was repeated throughout the participant districts' focus groups. Teachers in one district spoke of flexible definitions of what made for a "distinguished" vs. "proficient" rating:

> I think there's… even with schools if you have multiple administrators they have different definitions of basic versus proficient versus distinguished and it should be objective and clear cut because of the rubrics that are available."

> Some administrators feel that 'distinguished' is only a once in a lifetime type of reward to give a teacher and others feel 'basic' is the new normal.

These comments show teachers concerned about consistency in the ratings, and a link between the issue of reliability of the protocol itself –as teachers believed the protocol rubrics to have the potential for reliable use –and the less reliable subjectivity of the raters.

Some of the comments around reliability of scoring and the strictness of scoring each level, particularly the highest, could be read in multiple ways. As one teacher said,

> I actually had my [post-observation] conference with my principal at which point she said, 'If you earned [a distinguished rating] I'm going to give it to you,' but I know that's not happening in other places… For me here, my principal is a little more understanding to the fact that she hired us; she trusts us, so most of us should be getting distinguished."

The first part shows additional evidence that the scoring rubric is not inherently more objective and does not necessarily generate reliable scores, as one rater may be more willing to give out "distinguished" ratings than another. This appears to be the clear intent of the first part of this quote.

The second part is hazier. That "most of us should be getting distinguished" may simply reflect this teacher's belief that many of the teachers in her school are exceptional, which was a common thread in at least three participant districts. A more critical viewpoint would see this as evidence of insider bias, claiming that a principal is more likely to bend the observation protocol to the widget effect. This may influence teacher data use more toward the goal of scoring better by manipulating observers rather than for improved pedagogy. Reliability of the process was, overall, a significant question mark in the minds of teachers, but clearly a focal point of perceptions about the usefulness of feedback data.

Data Quality Summary

Though there was less direct data on teachers' perceptions about how data quality factors affected their use of data than system functionality, there are many indicators that these perceptions do play a role in shaping the data use process. The role of the evaluation protocol was not deeply explored, though teachers did note that the noise of protocol items they didn't feel was important to quality instruction was an issue of annoyance and thus potentially important in understanding how they continue to process and filter the feedback data they receive.

Validity and reliability of the data were of particular importance, though teachers varied on their perceptions of each. Teachers were generally balanced in surveys as to the question of whether or not the observation system was fair and accurate. While not psychometric validity, these give a benchmark to begin to understand teachers'

perspective related to validity. Teachers in focus groups typically found their experiences in observations to be fair, again suggesting that teachers perceived that the data was reasonably valid. This does not address the question as to its objective validity, but teachers at least were inclined to be comfortable with their results. Reliability was a much more a concern; teachers data use was likely negatively affected by their experiences of inconsistent, unreliable scoring.

CHAPTER 6: DISCUSSION

Feedback Filter: Implications of Teacher Data Use Perceptions

Teachers' perceptions about data use and factors affecting data use suggest several ways in which the findings may be applied to theory and practice. These findings show that teachers used, and failed to use, data in a number of ways. Further, teachers noted the effects on data use of several system functionality and data quality factors. From these findings, several theoretical and practical implications arise. This chapter will summarize these findings and outline the hypothetical links that can be drawn for teacher data use. Additionally, implications for evaluation policy and practice will be discussed. Finally, the limitations of the study will be addressed and suggestions for future lines of research posed.

Summary of Findings

The research questions addressed in this study were designed to understand teachers' perceived use of observational data. Teachers reported three different types of data use. These were changing practice to improve pedagogy, changing practice to generate higher scores on future evaluations, and non-use.

The primary formative purpose of teacher evaluation supposes that teachers will take the feedback provided to them and apply it to their practice in ways that will benefit student learning and achievement. Indeed, some teachers in this study did report that they had altered their teaching to improve their pedagogy. Keeping in mind that these data are self-reported, teachers improving practice is the ideal outcome of this type of policy

initiative. Though educational reform has been dotted with policy and reform efforts that have left little lasting impact over time (Olson, 2004; Tyack & Cuban, 1995) this exploratory finding is heartening for policymakers and those who continue to seek educational reform through changes in policy (e.g. United States Department of Education Office for Planning Evaluation and Policy Development, 2010).

This finding, however, was tempered by the somewhat more limited response from teachers in focus group interviews. While roughly 40% of survey respondents believed that they were changing their practice for the better based on data received from observations, this response was not as commonly heard during interviews. As with many teacher evaluation reforms currently taking place, there is only a small body of research that evaluates the effectiveness of these programs (e.g. Toch & Rothman, 2008). That teachers report changing practice for educational improvement is therefore a highlight, yet in need of further evaluation, as many other research-based intitatives (Slavin, 2002).

A second finding was that some teachers claimed to use observational data to learn about the particulars of the observation protocol. Some of this data use is to be expected, as teachers were experiencing a transition to new procedures and practices in their observations and wished to have a deeper understanding of the system in which they were now participating. This follows the current understanding of the process of educational change (Fullan, 2007). Additionally, some teachers noted they or others had used the data to learn how to "game the system" to achieve higher observation scores without real changes to their practice. Using knowledge of requirements of the current system to succeed has been noted in schools (Cullen & Reback, 2006) and by students (Baker, Walonoski, Heffernan, Roll, Corbett, & Koedinger, 2008), as well as by teachers

in other situations (Podgursky & Springer, 2007). Therefore this finding coincides with the current research, and suggests that there will be a segment of any population that will use characteristics of the system to improve scores rather than making more earnest changes.

This finding may not be wholly negative. One may argue that even changing for purposes of scoring may indeed benefit teaching, a "fake it until you make it" mentality. If one believes that the teaching behaviors in the protocol constitute a reasonable picture of effective teaching, then teachers who work to learn more about the requirements for compliance or to "game the system" may end up becoming stronger teachers despite themselves. This type of effect would require a concerted research effort to uncover.

A third and most prevalent of these variations was teachers reporting that they did not use the data provided to them. The finding that a sizeable minority of teachers –by choice or circumstance – report no use of observational data defeats the purpose of DDDM (Datnow, Park, & Kennedy-Lewis, 2012; Marsh, Pane, & Hamilton, 2006) and highlights a possible failure of the pilot program While potentially disheartening in terms of the formative purposes of teacher evaluation, looking at the various reasons that teachers fail to use the feedback provided to them from observation may point to ways in which the observation process can be made more useful, especially when combined with understanding how the implementation of the system may affect teachers.

One common reason cited by teachers for not using observational data was that teachers simply did not find the data given them useful. While studies in teacher data use and DDDM account for several of the factors that influence data use, there is less focus when teachers do not use provided data (Ikemoto & Marsh, 2007; Marsh, 2012). Though

this study found a sizable number of teachers not using data, more research is needed into directly addressing this finding. This study looked at several factors that potentially affect teacher data use in the context of evaluative observations in order to develop this picture.

Factors affecting data use

System Functionality. Several factors were noted in Chapter 5 as relating to teachers' ability to use or disuse data. System functionality issues were most common for teachers to note. Among these, the capacity of the observer was most predominant. This finding is complementary to Ikemoto and Marsh (2007), among others (see, e.g. Dembrosky, Pane, Barney, & Christina, 2005; Mason, 2002) who found that teachers themselves experienced capacity gaps when attempting to use data. In this study, teachers noted that the capacity of the observer was also a necessary factor in promoting their data use. Teachers very much believed that observers who did not know the content area of the observed teachers were less likely to give useful feedback. The same was true of observers who were perceived as being unfamiliar with the observed teacher's class or school. Teachers felt strongly that without deep knowledge of the classroom context, the feedback received would be limited in its usability. These findings support those laid out by researchers noting the importance of observer capacity (Knapp, Copland, & Talbert, 2003; Park & Datnow, 2009; Prestine & Nelson, 2005). This sets up a conundrum: it seems unlikely that in a given district there would be administrators familiar with each and every content specialty and grade level taught, yet teachers were strongly opposed in most cases to the use of external observers who might be more familiar with their content if they were not seen as being connected with the context of the classroom. This finding suggests both theoretical and practical considerations, discussed below.

In addition to content and context, teachers believed that time devoted to observation limited their data use, a finding in line with other data use researchers (Ikemoto & Marsh, 2007; Wahlstrom & Louis, 2008). When provided too little time to engage with and process observational data, teachers did not feel they could fully benefit from the data received, a finding 30 years removed, yet similar to the time pressure noted by Huberman (1983). As teachers progressed in the system, time seemed to be less of an issue. This suggests that teachers experience with the observation process and protocol may allow data use to grow given a long enough implementation. This may be tied to teachers' use of data to understand the system. As they grow in the knowledge of what is expected of them, the time used to interact with the data can be more focused on use for other purposes.

Closely related to time, the logistics and schedule of the observation also seemed to have an effect on teachers' perceptions of data use. Most important to teachers along these lines was access to the data in a reasonable time frame. As time between the observation and feedback received stretched to three weeks or more, teachers frequently reported that the data had lost some usability due to this lag. Their reason was simple. As time passed, teachers were less able to recall the specifics of their observed lesson and thus could not apply feedback as readily as the context had faded. Once again, teachers' perceptions of the important factors influencing their data use align with previous findings in terms of evaluation in general (e.g. Peterson, 2000; Weisberg, Sexton, Mulhern, & Keeling, 2009).

Finally, trust between teachers and the system was a potential barrier to data use. Though teachers mainly believed that the purposes of the system were formative for them

–though whose responsibility formative development was unclear, as noted above –and did not fear severe consequences such as tenure loss as a significant threat, teachers were very aware of summative measures in place for the evaluation system. Teachers felt threatened in general by the summative aspects they saw in place, which may have affected their use of the data, focusing them on scoring well to avoid punishment rather than using data to improve their overall pedagogy. Here, too, the findings are complementary to research in other educational contexts. As Louis (2007) showed that trust was essential to the success of school reform initiatives, it also appears to be important to teacher data use. It is reasonable to hypothesize from this finding that high-trust schools and teachers who trust the purposes of the system are more likely to use data from observations.

System functionality must be addressed in order to understand the ways in which teachers use observational data. In each of these factors, teachers showed that system functionality was a primary force affecting their use of evaluation data. The alignment of these factors, or working with teachers to address system functionality will likely help to create more frequent and effective teacher use of observational data.

Data Quality. Beyond system function, data quality was another category that teachers found to affect their data use. Teachers' opinions on the usefulness of the protocol, and the specificity of the data provided were cited as influential in data use, though to less extent that functionality factors. In short, teachers believed that specific feedback was more helpful to their ability to use data than more general information. This was one way in which the new protocols were successful in promoting data use. Teachers were mixed, however, on whether they agreed that the protocols themselves measured

what they believed was important to teaching and could be used to assess their performance appropriately. This finding is in keeping with research that suggests that there is not a unified view of what constitutes effective teaching (Stigler & Hiebert, 2009) or that there is a difference between quality teachers and quality teaching (Bell, Gitomer, McCaffrey, Hamre, Pianata, & Qi, 2012). The issue of protocol relates directly to how teachers use data, as if they do not believe that the protocols are measuring useful teaching behaviors they may be less likely to find the data valuable and, therefore, less likely to engage in any form of data use, as noted above.

Teachers also were mixed on their feelings about data validity and reliability coming from the pilot. Roughly the same percentage of teachers found the observation protocol fair and unfair; accurate and inaccurate. Though teachers didn't address objective, psychometric validity issues, it seems that they were split on whether they believed the results to be valid. In focus groups, however, teachers seemed somewhat more content with the issue of validity; teachers did not voice strong beliefs that their results were out of line with where they believed themselves to be as teachers. This finding aligns with studies in data use in general (Ikemoto & Marsh, 2007), but would likely raise concern with advocates of psychometric validity (e.g. Bell, Gitomer, McCaffrey, Hamre, Pianata, & Qi, 2012; Braun, 2005) and experimental research such as Slavin (2002). Though a factor in encouraging data use, teachers do not seem as intimately concerned with the psychometric characteristics of the data as they do with issues of system functionality.

Reliability was largely an issue of inter-rater reliability. Teachers routinely noted that their scores differed depending on their observer and thus were less inclined to take a

given set of feedback and use it confidently. Reliability concerns are an issue in a number of studies (e.g. Brandt, Mathers, Oliva, Brown-Sims, & Hess, 2007; Cash, Hamre, Pianata, & Myers, 2012; Darling-Hammond, Amrein-Beardsley, Haertel, & Rothstein, 2011); teachers, too, tended to find unreliable ratings troubling to their ability to use data. This finding also relates to the issue of trust; reliable results may encourage the overall feeling that the system is trustworthy. Reliability is already a concern of the pilot program and many other current efforts at evaluation reform; it comes as little surprise that teachers here identified reliability as an issue affecting data use.

Overall, though issues of data quality were noted by teachers as having an effect on their ability to use data, they seemed to be less prevalent than issues of system functionality. These factors contribute to the way teachers filter data, decide whether it is useful, and ultimately, how to use it, but are secondary to their understanding of whether the data will be functional for them as the program is implemented.

Contribution to the Literature

The primary contribution of this study to the literature around teacher data use is expanding the field into the area of teacher data use in evaluation. The mixed-methodology approach used for this study allowed for both a higher-level understanding of teacher data use as well as the factors that affect it, as well as a qualitative analysis of these factors that suggest specific ways that these factors facilitate or hamper data use. The findings confirm many previous studies that suggest that the factors outlined in the conceptual framework are indeed important to teachers as they decide whether to interact with observational data or filter this feedback into non-use. Thus the conceptual framework developed for this study appears to be a starting point for more focused study

into how teacher data use in this context works. This opens up the unknown process shown in the skeleton framework (Figure 1) and provides a tentative model to be used to explore these findings more deeply.

Theoretical Implications

In all, these findings suggest a number of theoretical implications related to teacher data use in terms of the theoretical model proposed in this study. The data-use factors explored here seem to be used by teachers to create a filter for the feedback received from their observations. This "feedback filter" may all but eliminate data use for teachers. If too many of the factors explored here are set to minimize usability in teachers' minds, data will simply be ignored. However, the filtering process that these findings suggest may also help to focus teachers on areas that may most impact teaching, provided that the filter "settings" are correct. In order to understand how teachers use data and maximize use toward the desired outcomes of improved teaching practice and student learning, it is essential to learn more about how this feedback filter functions for teachers. By understanding the filter, one may be able to engineer conditions that help teachers to focus on what is important in improving their practice while eliminating noise in the system.

This study set out to develop hypotheses about how system factors may affect teacher data use in evaluation. As noted, these factors formed a sort of feedback filter that, in general, served as a way for teachers to sift through the feedback provided, sometimes resulting in total disuse of the data. There are several ways in which these findings suggest how the feedback filter is constructed and may be manipulated; three such hypotheses are highlighted here.

Observer Capacity as a Key Data Use Lever

Teacher's perceptions of the capacity of their observer were a key issue in teachers' use of the data. This has been noted in other data use contexts in education (Ikemoto & Marsh, 2007), and can be extrapolated to the context of observational data use. When teachers did not feel that their observer was qualified to give high-quality, feedback, their willingness to interact with and use data was severely limited. This suggests that the content and pedagogical content knowledge of the observer, as well as contextual familiarity with the observed classroom are central to teacher data use. This further suggests that data use is linked to teachers' perceived level of observer capacity; the greater the perceived capacity, the more likely teachers are to allow data to pass through the filter.

Observer capacity is linked to other factors, as well. Teachers believe that feedback given by low-capacity observers is less specific; data specificity is a factor linked in teachers' minds to greater data use. It is possible that an observer that teachers do not view as in tune with their content and context may be able to provide data that reaches teachers' desired level of specificity, but this seems unlikely as they may be predisposed in this scenario to minimize the value of the feedback. It is therefore hypothesized that the observer's capacity may be more influential in teacher data use than specificity.

Additionally, teacher perceptions around observer capacity seem likely to affect perceptions of validity, but be dependent on perceptions of reliability. If teachers feel that their observer is of high capacity, they may be more likely to view the overall observation as valid, further promoting teacher data use. In the case where a teacher already believes

the protocol to be invalid, a strong observer may influence teacher perceptions by providing useful feedback. Reliability, on the other hand, will likely influence teacher perceptions of capacity. If raters are reliable, teachers are more apt to see capacity perceptions affected, or the role of their perceptions about capacity diminished. If three observers all see their practice the same way, then perhaps there is something to the feedback that demands use.

In these ways, observer capacity, especially the content knowledge and contextual familiarity of the observer, seem to be important levers in affecting teacher data use. If these two factors are aligned in teachers' minds, there is less chance for teacher filtering of the observational data, which in turn promotes greater data use. This suggests at least one pathway by which the veil of data use may be pierced in the future. Yet this lever is potentially stymied by the makeup of observers used; a practical consideration discussed below.

Trust in the System Promotes Data Use

Trust is a factor that reaches across many of the factors explored in this study. Teachers mainly responded with trust issues stemming from the purposes of the system; greater distrust may lead to less use of data. There are two ways to hypothesize the effects of this kind of trust for data use. First, in schools and districts where teachers have a greater comfort in the formative purpose and less concern with the summative aspects of evaluation, it is likely that teachers will more often use data formatively, that is, to change their teaching to better reflect the feedback they receive. The hypothetical ideal of this type of trust in an evaluation system without summative purpose, solely designed for

teacher development. Ironically, this is how the observation protocols in this study were designed; they were employed both formatively and summatively in the pilot.

The converse situation may hold true as well. In situations where the purpose is perceived as more summative there would likely be less data use. Also, data use may be more focused on scoring well versus teacher growth for improved pedagogy. In either case, trust in the purposes of the system will hypothetically affect teacher data use.

Further, trust in the system may extend to trust in other areas. As teacher trust in the system increased, their perceptions around validity and reliability may increase. Noting that this perception may not be related to objective measures of validity and reliability, teachers who see observation data use valid and reliable are likely more inclined to use such data. Trust may play a role in perceptions of observer capacity, as well. For example, if an external and internal observer are both perceived to have equal capacity for content knowledge in an observation, greater trust –in this case likely to be bestowed upon the known observer –may color the teachers' perceptions of capacity in the observer and, consequently, their use of the data. Trust, therefore, is also an important factor in considering how teachers use observational data and should be acknowledged when endeavoring to understand the makeup of teachers' feedback filter.

Low Quality Data Threatens Data Use

One of the most distressing findings of the study was that a significant portion of teachers were reporting no change to their practice due to the observation data received. Since one goal of heightening teacher evaluation is to encourage teachers to use feedback to improve their teaching, teachers who do not use the data are a problematic side effect. One of the main reasons posed by teachers for not using data was that they did not find it

useful to their practice. This suggests that the factors posed in the conceptual framework are especially important in view of how to increase the perceived utility of the data. Teachers seeing data as useless turns a feedback filter –a screening device the eliminate data perceived as unhelpful –into a solid wall, removing the opportunity for even meaningful data to be used.

Teachers' failure use data may suggest larger issues, as well. If teachers see data as useless, then it becomes important to understand the genesis of these perceptions. This may point to a systemic issue in terms of digging into the interplay of the factors outlined in this study. The block between factors that may affect data use and the process of using that data is an important idea and points to a need to bring more depth into the understanding teaches possess of these factors, learning how to both create data of sufficient usefulness for teachers and making sure that the system of observation and data delivery facilitates use. This implies that there are, indeed, a number of avenues of future research to expand the theoretical framework; this issue is discussed below. Moreover, there are a number of implications for the practice of implementing a new evaluation system in schools.

Implications for Policy and Practice

As the theoretical model for understanding teacher data use in evaluation continues to be filled out, there is an immediate need to put that which has been learned to practical purposes. Since so many states, districts, and schools are currently focused on restructuring their teacher evaluation programs, the time is ripe to put knowledge into practice. Though there will undoubtedly be multiple evaluation program iterations –there

have been, already –getting knowledge into practice may help those teachers currently experiencing these new programs.

Using External Observers and Observer Capacity

Observer capacity, particularly in the domains of content knowledge and contextual familiarity, was a significant data use effector, according to teachers. It will likely be difficult, however, to find individuals who possess both robust content expertise as well as intimate contextual knowledge of the observed classroom. Added to this is the issue of validity. Past research has suggested that external observers, with no relationship to the school or teacher observed, are likely to have more valid ratings (Graham, Milanowski, & Miller, 2012; Little, Goe, & Bell, 2009). Even within this argument, there are certain levels of nuance. While external observers may not be as likely to experience the specific biases of principals or other insiders, having less internal pressure to create high scores (Weisberg, Sexton, Mulhern, & Keeling, 2009), the biases or misunderstandings that may arise from limited contextual knowledge of external observers are not well explored. This creates a conundrum of how to balance valid ratings with encouraging teachers to use observational data.

The findings here make clear that teachers prefer a direct supervisor in terms of comfort in using the data and perceiving the results of their observations as useful. It is likely that teachers will continue to experience supervisors as observers. Whether due to the difficulty in finding trained external observers or the costs of hiring such observers, principals are likely to remain primary observers. This largely ameliorates teacher concerns about contextual familiarity, but still falls short on content knowledge. After all, there is only so much a given principal can master in terms of subject area expertise.

Here, external observers might play a positive role. External observers, when available, could be selected to match with content expertise; less common subject area experts, such as music or art, could be used across multiple schools or districts. This would bring the benefits of neutral observation as well as content expertise. As external observers are already becoming a common feature of enhanced observation, tailoring these observers to subject areas could be an added requirement. There would, of course, be potential difficulties in developing the correct pool of content experts. Nevertheless, it may be possible with determination.

Teachers, however, as shown in this study, struggle to trust external observers. Building this trust will likely be a key aspect of implementation in a system that attempts this type of distributed observer capacity. This challenge may be met by helping teachers to understand the benefits of content-specialized observation, even by an outsider, and by using the same external observer for multiple observation sessions. This may increase teacher trust in these ratings, leading to a higher potential for data use.

Make Logistics and Scheduling a Priority

A simpler practical implication of this study to accomplish may be a focus on the logistics and scheduling of the observations and post-observation conferences. Teachers felt that delays in both the observed lesson itself, and, more importantly, receiving feedback, greatly reduced the usability of the data. Therefore, schools and districts should make scheduling and keeping the schedule a priority in order to foster greater data use for teachers.

A primary challenge of trying to prioritize these issues is found in both time and labor costs. Increasing the number of scheduled observations should increase the

accuracy, reliability and validity of ratings, but with increased cost in time and labor. These barriers must be balanced against the scheduling and logistical issues necessary to create an observational pattern that promotes data use and improvement of practice.

Naturally, this may be difficult in the case of using principals as observers. There will always be emergencies and interruptions to the principal's schedule given the number and nature of responsibilities of school administrators. When possible, however, observation sessions should be considered a top priority; more so for rapid turnaround and delivery of feedback. Larger districts can accomplish this by assigning other administrators to be responsible for certain aspects of school management during observation and post-observation conference periods. Smaller districts may struggle more with this issue as they will have less personnel to cover emergent problems.

In the case of external observers, there will likely be less issue of holding to the schedule, as they would likely have a smaller number of other responsibilities as a principal. Of course, this will likely be a challenge in places like the pilot program studied here, where many districts chose to use central office administrators or even principals from other district schools, as external observers. These people could prove more difficult to keep to a schedule and prompt feedback. In the event that external observers were content matched and their duties more limited to the observation process, there may be more success.

In all cases, getting data to teachers in a reasonable time will most likely encourage a greater use of the feedback data.

Creating Ownership and Encouraging Data Use

Teacher perceptions that the feedback received is useless may be a controlling barrier to data use, an, as such, should be addressed and avoided. Teachers should be encouraged to take ownership of their own development. This is already in place in some systems, where teachers are required to create an individualized professional development plan (e.g. Williams, 2009); the findings here suggest that those should be taken seriously and teachers given assistance in realizing those plans on an individual level. This will likely be a challenge in many districts, as the culture of school-wide, corporate professional development is firmly entrenched. Yet bringing professional development back to individual teachers may be a fundamental issue in promoting data use.

Additionally, special attention should be paid to creating a system that helps to open teacher feedback filters to the data provided in observations. This may include, once again, the use of both internal and external observers along with a process to help teachers grow to trust these observers. This could perhaps be via repeated observations by the same external observer and reducing the perception that summative sanctions are more a key purpose of the system; focus on the formative aspect of observation may also increase trust. Though more study is needed to develop the ways in which the various factors affect one another, understanding the workings of the filter may help to increase teacher perceptions that the feedback data is useful.

If teachers see the data as useless, the feedback filter is closed. In many ways, without teacher ownership of their development, and encouraging their interaction with the data in terms of creating data that teachers see as useful, other systemic features and

changes may be moot. Teachers must first understand the reasons that they *should* interact with the observational data before moving on to helping them to do so more effectively. Though a difficult challenge to overcome, understanding and dealing with teacher passivity is an important issue to pursue.

Study Limitations

This study faced a number of limitations that should be considered when considering both the findings and implications presented herein. First, the study was conducted in the early stages of a pilot program, and teachers were just getting used to the protocols and procedures used in the new evaluation program. In Year 1 of the pilot, some districts found out after the beginning of the school year that they had received pilot funding. This compressed the time available for districts to train teachers in the observation system and to conduct the required observations. Therefore, as noted above, these findings should be considered tentative, as teachers' perceptions of their data use may evolve with multiple years of experience with the protocol. Confirming the findings of this study by revisiting participating districts after a few years of full implementation would help to reduce this limitation.

Another primary limitation is with the sample of schools participating. Each district was self-selected; they sought after and received funding from the state to be a part of the pilot program. Therefore the sample surveyed and interviewed cannot be considered fully representative of teachers in this situation. In both the survey and focus groups, there are additional limitations with respect to representativeness of the sample. In the survey, the response rates were lower than may be considered optimal, with 59% and 39% responding in Years 1 and 2, respectively. Teachers who chose to participate in

the survey self-selected, and may not be a random, representative sample. Focus groups, too, were not randomly conducted. Teachers were selected by the project directors in each district rather than drawn randomly from the available pool.

Additionally, there were a number of variations involved in the implementation of the pilot programs across districts. Schools were given a relatively large amount of freedom in choosing how they would conduct observations –though within the bounds of their selected protocol –and who would conduct the observations. Thus the procedure was not experimentally controlled in this way. This would be a serious barrier to the study if its purpose were to generalize findings; the purpose of this study was primarily to generate hypotheses for further understanding of the theoretical framework and thus these limitations may not cause grievous problems in the understanding of the findings.

A second limitation of the study was that it draws data from a larger study, one whose purpose was not solely to understand teacher data use. While there were significant questions and commentary across both the survey and interviews regarding how teachers were using or not using observational data, the overall scope of the larger research was to analyze and assess the pilot program itself. Therefore this work constitutes only a segment of the collected data and is a secondary analysis of that data. Once again, this study has attempted to generate only tentative findings rather than fully generalizable absolutes, and thus may not suffer severely from this limitation.

A final limitation to the study lies in the collected qualitative data. Since the data from focus groups was collected in two different ways in Years 1 and 2, the analysis was not conducted the same way across both years. In Year 1, interviews were recorded and transcribed, allowing more direct use of quotes and coding. In Year 2, the development of

the site visit guide and report led to a more developed thematic picture of visited districts, but made for a higher level of analysis. There were fewer direct quotes available, and the sentiments were filtered through the overall understandings of the site visit team. While codes and themes did transfer between transcripts from Year 1 and site visit guides in Year 2, the data from Year 2 is less focused than Year 1 transcripts. This limitation should be noted, but likely does not affect the findings outlined within this study.

By and large, the findings presented here do seem to suggest that teacher data use may be affected by how teachers see the various factors explored here should, even limited by the above, begin to paint a picture of how the theoretical framework may be linked across teacher data use and, in many ways, expanded beyond the bounds of direct contributing factors. Further research on a large scale may help to confirm the overall generalizability of these findings.

Avenues for Future Research

Spurring these future research effects was the primary goal of this work. Naturally, then, the hypothetical links uncovered by the analysis of this study should be a rich vein of future research efforts. Future work in this field would involve more controlled, experimental testing of the links between the factors affecting teacher data use explored here and how these factors encourage, discourage, or change teachers' proclivities toward using observational data. Additionally, this study suggested that there were links between the factors themselves, another area of research that could be fruitful. Each of these avenues of inquiry has the potential to strengthen the understanding of how teachers use data from observation. This may aid in understanding the makeup and

function of teachers' feedback filter, giving a greater knowledge of how to truly understand and encourage teacher data use.

Further, research may be undertaken to understand not only how these factors relate to teacher data use as a whole, but how each affects *meaningful* data use. This requires a more tightly conceptualized definition of what this looks like in practice. One simple definition would be that of teachers using data to change their practice toward behaviors that have been shown to be more effective for student learning. Therefore not only could the affective factors be studied for how they relate to data use as a whole, but how each changes data use in a more desirable way. In both these cases, future research is needed to confirm the tentative hypotheses developed here. By implementing these ideas in a larger design, the overall strength of the hypotheses could be tested and confirmed.

These two paths of further research would help to deepen understandings of how the theoretical framework for teacher observational data use developed for this study functions within the context of application in schools. Yet the findings around non-use of feedback suggest that there are further layers of the framework to be explored and uncovered. The theoretical framework becomes, in some ways, like reverse engineering an onion: adding layers until a greater understanding of the whole is more complete. No initiative or reform in education happens in a vacuum; further research into how teacher observational data use is affected by the outer layers of the framework is needed. Indeed, understanding what those layers are and the relative placement of teacher observational data use in the overall landscape of education research is an area that is ripe for inquiry. Given that teacher data use itself is a relatively young field, and much of this learning has

not been focused on its use in observation, the overall field of this type of research is ready for a greater focus in the future.

Conclusion

Probing the form and function of the feedback filter in teacher evaluation data use requires a great deal of consideration. It seems apparent that teachers develop an internal system of feedback filters that may help or discourage data from being used based on a number of potential factors. Observer capacity, trust in the system and the people in it, the validity and reliability of data, and the observation protocol itself, among others, contribute to the web of interconnected issues influencing teachers' willingness and ability to use data received in observations. Beyond this, the way teachers approach this data use is also complex. Teachers may use the data to improve their practice. They may use the data to preserve their summative scores and achieve higher ratings in the future, even when they mainly conceive of the evaluation as designed to promote their formative development. And they may not use the data at all, the feedback filter opaque to the data received, or even see the data as an object of use at all.

In all these ways, teacher data use in observation is an area that will play a role in how evaluation reform efforts are undertaken across the country, and will likely contribute to the perceived –or even real –success or failure of these initiatives. By beginning to understand the possibilities of how the data use process works for teachers in this arena there is an opportunity to get at the goals of evaluation in less of a black box manner than previously attempted. Though there will be many other initiatives competing for the attention of teachers, schools, districts and even states, teacher evaluation will

undoubtedly be a constant area of effort and should therefore benefit from a greater depth of knowledge about its workings. Rather than simply hoping that positive, intended outcomes from observation will occur, working to understand how to encourage those outcomes will likely result in a more fruitful effort. This study may help to contribute to that conscientious movement toward a better understanding of the process and, ultimately, a more effective teacher evaluation program.

APPENDICES

Appendix A: Year 1 Teacher Survey

Professional Background

Q1 Are you a teacher currently teaching students?

- ❍ Yes
- ❍ No

Q3 What is the highest degree you have earned?

- ❍ Bachelor's degree (B.A., B.S., etc.)
- ❍ Master's degree (M.A., M.A.T., M.B.A., M.Ed., M.S., etc.)
- ❍ Educational specialist or professional diploma (at least one year beyond master's level)
- ❍ Doctorate or first professional degree (Ph.D., Ed.D., M.D., L.L.B., J.D., D.D.S.)

Q4 Which grades do you teach? Select all that apply.

- ❑ Pre-K
- ❑ Kindergarten
- ❑ Grade 1
- ❑ Grade 2
- ❑ Grade 3
- ❑ Grade 4
- ❑ Grade 5
- ❑ Grade 6
- ❑ Grade 7
- ❑ Grade 8
- ❑ Grade 9
- ❑ Grade 10
- ❑ Grade 11
- ❑ Grade 12

Q5 Which statement best describes the way YOUR classes at your current school are organized?

- ❍ You instruct several classes of different students most or all of the day in one or more subjects (sometimes called Departmentalized Instruction).
- ❍ You are an elementary school teacher who teaches only one subject to different classes of students (sometimes called an Elementary Subject Specialist).
- ❍ You instruct the same group of students all or most of the day in multiple subjects (sometimes called a Self-Contained Class).
- ❍ You are one of two or more teachers, in the same class, at the same time, and are jointly responsible for teaching the same group of students all or most of the day (sometimes called Team Teaching).
- ❍ You instruct a small number of selected students released from or in their regular classes in specific skills or to address specific needs (sometimes called a "Pull-Out" Class or "Push-In" Instruction).

Q6 What is/are the subjects you currently teach? Select all that apply.

- ❑ 21st Century Life and Careers
- ❑ English Language Learners (ELL)/English as a Second Language (ESL)
- ❑ Health and Physical Education
- ❑ Language Arts
- ❑ Mathematics
- ❑ Science
- ❑ Social Studies
- ❑ Special Education
- ❑ Technology
- ❑ Visual and Performing Arts
- ❑ World Language
- ❑ Other: _____

Q7 How many years will you have been teaching at the end of the current school year?

- ❍ 1-3
- ❍ 4-6
- ❍ 7-10
- ❍ 11 or more

Current Teacher Evaluation in Comparison to Previous System

Q8 Please indicate how much you agree or disagree with the following statements about teacher evaluation systems in general.

	Strongly agree	Agree	Neither agree nor disagree	Disagree	Strongly disagree
Teacher evaluation is essential to raise the standards of teaching and learning.	❍	❍	❍	❍	❍
Teacher evaluation should primarily focus on the identification of my professional development needs.	❍	❍	❍	❍	❍
Teacher evaluation aims at meeting the minimum standards.	❍	❍	❍	❍	❍
Teacher evaluation aims at providing useful information for teachers to improve their performance.	❍	❍	❍	❍	❍
Teacher evaluation should be based upon a list of professional	❍	❍	❍	❍	❍

competencies or behaviors.					
As a professional, I am entitled to have my performance appraised.	❍	❍	❍	❍	❍
Teacher evaluation should aim primarily at making managerial decisions.	❍	❍	❍	❍	❍
Teacher evaluation aims to enhance teachers' reflection on their practice.	❍	❍	❍	❍	❍
Teacher evaluation should be used both for professional development and accountability purposes.	❍	❍	❍	❍	❍

Q9 In comparison to your previous teacher observation system, how would you rate the current (new) teacher observation system on the following dimensions:

	The current system is much better than the previous system	The current system is better than the previous system	The current system is neither better nor worse than the previous system	The current system is worse than the previous system	The current system is much worse than the previous system	Does not apply	Don't know
Formalization (clear rules, steps, procedures, reporting forms)	❍	❍	❍	❍	❍	❍	❍
Ease of use	❍	❍	❍	❍	❍	❍	❍
Grounding in research	❍	❍	❍	❍	❍	❍	❍
Intuitiveness	❍	❍	❍	❍	❍	❍	❍
Usefulness for providing guidance to teachers	❍	❍	❍	❍	❍	❍	❍

Q10 Were you responsible for informing other teachers about the new teacher evaluation system?

❍ Yes
❍ No

Training on the New Teacher Evaluation System

The next set of questions is about the training you received on the new teacher evaluation system in your district as part of the pilot program. There are no right or wrong answers. We are simply interested to learn about your experience with the new teacher evaluation system.

Q11 How many hours of training or education have you personally received on the new teacher observation system:

- ❍ 0
- ❍ 1-2
- ❍ 3-4
- ❍ 5-8
- ❍ 9-16
- ❍ 17-24
- ❍ 25-32
- ❍ 33-40
- ❍ More than 40

Q12 Overall, how well, would you say, the training accomplished each of the following:

	Very well accomplished	Accomplished	Somewhat accomplished	Not accomplished	Not at all accomplished	Does not apply
Help you understand your district's system of assessing teachers	❍	❍	❍	❍	❍	❍
Help you understand the main components of the teacher evaluation: teacher practice and direct measures	❍	❍	❍	❍	❍	❍

of student achievement						
Help you understand the process of linking student growth scores to teacher observations in tested subjects	❍	❍	❍	❍	❍	❍
Help you understand the process of linking student growth scores to teacher observations in non-tested subjects	❍	❍	❍	❍	❍	❍
Help you understand the information needed for you to be accurately assessed	❍	❍	❍	❍	❍	❍
Help you understand the criteria for assessment of teachers' planning process	❍	❍	❍	❍	❍	❍
Help you to understand the criteria for	❍	❍	❍	❍	❍	❍

assessment of teachers' instructional practices						
Help you to understand the feedback after an observation	❍	❍	❍	❍	❍	❍
Help you to understand what underlies judgments of teacher quality	❍	❍	❍	❍	❍	❍
Help you to understand potential biases in the way teachers are evaluated	❍	❍	❍	❍	❍	❍

Q13 Have you been evaluated by the new teacher observation system?

❍ Yes, I have been evaluated as part of the new evaluation system at least once.
❍ No, but I will be evaluated in the future.
❍ No, and I will not be evaluated.

Evaluation of New Teacher Evaluation System

This question asks for your personal evaluation of the new teacher evaluation system based on your experience.

Q14 Below is a series of statements about the new teacher evaluation system used in your school district. For each statement, please indicate whether you strongly agree, somewhat agree, neither agree nor disagree, somewhat disagree or strongly disagree with that statement. Please answer based on your personal experience and observation. Remember that your answers are confidential.

	Strongly agree	Agree	Neither agree nor disagree	Disagree	Strongly Disagree	Does not apply	Don't know
I feel comfortable being assessed by the district's new evaluation system.	❍	❍	❍	❍	❍	❍	❍
The district's system for assessing teachers generates accurate assessments.	❍	❍	❍	❍	❍	❍	❍
The district's system for assessing teachers is fair.	❍	❍	❍	❍	❍	❍	❍
The district's system for assessing teachers generates assessments that provide constructive individual feedback and promote professional development.	❍	❍	❍	❍	❍	❍	❍
The district's	❍	❍	❍	❍	❍	❍	❍

system for assessing teachers is well aligned with the district's curriculum.							
The district's system for assessing teachers clearly separates accomplished from unaccomplished teachers.	❍	❍	❍	❍	❍	❍	❍
The district's system for assessing teachers fits well with other school/district initiatives.	❍	❍	❍	❍	❍	❍	❍
The district's system for assessing teachers provides a firm basis for making teacher tenure and promotion decisions and weeding out weak teachers.	❍	❍	❍	❍	❍	❍	❍
The district's system for assessing teachers helps this district meet its accountability requirements under NCLB and other external mandates.	❍	❍	❍	❍	❍	❍	❍
The district's system for	❍	❍	❍	❍	❍	❍	❍

assessing teachers helps improve student achievement.							
The district's system for assessing teachers consumes resources that could be better spent elsewhere.	❍	❍	❍	❍	❍	❍	❍
The district's system for assessing teachers is relevant for my subject area and teaching methodology	❍	❍	❍	❍	❍	❍	❍

Q15 Please indicate how much you agree or disagree with the following statements about your perceptions of the new teacher observation system.

	Strongly agree	Agree	Neither agree nor disagree	Disagree	Strongly disagree
I feel adequately informed about the new observation system	❍	❍	❍	❍	❍
I feel that the new observation system takes too much time	❍	❍	❍	❍	❍
I do not feel prepared for the new observation system	❍	❍	❍	❍	❍
I understand the new observation system	❍	❍	❍	❍	❍
I receive useful feedback from observers under the new observation system	❍	❍	❍	❍	❍
The new observation system provides a fair picture of my teaching	❍	❍	❍	❍	❍

Q16 Please indicate how much you agree or disagree with the following statements about your experience with the new teacher evaluation system.

	Strongly agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
In my school evaluation criteria and indicators are appropriate.	❍	❍	❍	❍	❍
Existing instruments for teacher performance evaluation are clear.	❍	❍	❍	❍	❍
Existing evaluation criteria take into account the context of teaching.	❍	❍	❍	❍	❍
The evaluation process at my school allows teachers to explain decisions and actions.	❍	❍	❍	❍	❍
Rating scales used to evaluate my performance are appropriate.	❍	❍	❍	❍	❍
I am given useful feedback by the evaluator.	❍	❍	❍	❍	❍
I feel that in my school teachers' work and	❍	❍	❍	❍	❍

achievements are recognized.					
I feel that the evaluators in my school have the required knowledge and competencies to appraise teachers.	❍	❍	❍	❍	❍
I feel that the evaluators in my school have received adequate training to perform their job.	❍	❍	❍	❍	❍
In general, I think that the feedback that I am given focuses upon suggestions for improvement.	❍	❍	❍	❍	❍

Q17 Please indicate how much you agree or disagree with the following statements about your perceptions of the effects of the new teacher evaluation system.

	Strongly agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
The evaluation system encourages me to reflect on my teaching.	❍	❍	❍	❍	❍
The evaluation system has made me more aware of my strengths and weaknesses as a teacher.	❍	❍	❍	❍	❍
The evaluation system has led to an intensification of my work.	❍	❍	❍	❍	❍
The evaluation system has increased the bureaucratic work at school.	❍	❍	❍	❍	❍
The evaluation system has led to tensions among staff.	❍	❍	❍	❍	❍

Q18 I would prefer to be evaluated by:

- ❍ A superior (principal, etc.) with whom I have a developed professional relationship
- ❍ A superior with whom I am not very familiar
- ❍ A teacher with whom I have a developed professional relationship
- ❍ A teacher with whom I am not very familiar
- ❍ Someone whom I have never met

Q19 Why do you prefer to be evaluated by that person?

Q20 Which of the below BEST describes the relationship you have with the people who observe you

- ❍ I know them and they know me/my classroom well
- ❍ I know them and they know me/my classroom a little
- ❍ I know who they are but have no relationship with them
- ❍ I do not know them

Q21 Please indicate how much you agree or disagree with the following statements about your perceptions of how you would prefer to be evaluated in the new teacher evaluation system.

	Strongly Agree	Agree	Neither Agree Nor Disagree	Disagree	Strongly Disagree
I will excel under the new evaluation system	❍	❍	❍	❍	❍
I am confident that I will be accurately evaluated in the new system	❍	❍	❍	❍	❍
I feel comfortable being observed and evaluated by the current person responsible for it	❍	❍	❍	❍	❍
I am confident I would score well on an evaluation done by my principal	❍	❍	❍	❍	❍
I am confident I would score well on an evaluation done by an impartial observer	❍	❍	❍	❍	❍
I am more likely to be accurately assessed by	❍	❍	❍	❍	❍

someone who knows my classroom and teaching well					
Impartial observers will not understand the context of my classroom	❍	❍	❍	❍	❍
An impartial observer may give a more accurate evaluation of my teaching than someone who knows me	❍	❍	❍	❍	❍
I would rather be evaluated by a direct superior (i.e. a principal) than a peer or master teacher in my content area for purposes of accountability	❍	❍	❍	❍	❍
I would rather be evaluated by a direct superior (i.e. a principal) than a peer or master teacher in my content area for purposes of professional development	❍	❍	❍	❍	❍
I would rather be evaluated	❍	❍	❍	❍	❍

by an impartial observer than someone who knows me for purposes of accountability					
I would rather be evaluated by an impartial observer than someone who knows me for purposes of professional development	❍	❍	❍	❍	❍

Q22 Please indicate how much you agree or disagree with the following statements about your perceptions of how content knowledge affects evaluation.

	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree	Not Applicable
Evaluating good teaching in my subject area is different from evaluating good teaching in other subject areas.	❍	❍	❍	❍	❍	❍
A strong understanding of the pedagogy specific to my subject matter (i.e. the pedagogy of science or special education) on the part of the observer is essential for an accurate observation of my teaching	❍	❍	❍	❍	❍	❍
The person who evaluates me has a robust knowledge of the content I teach	❍	❍	❍	❍	❍	❍
The person who evaluates me has a robust understanding of what good	❍	❍	❍	❍	❍	❍

teaching looks like in my subject area						
I would prefer to be evaluated by someone who understands my content area deeply	❍	❍	❍	❍	❍	❍
I would be more accurately evaluated by someone who understands my content area deeply	❍	❍	❍	❍	❍	❍
The people who evaluate me do not understand the intricacies of teaching my subject	❍	❍	❍	❍	❍	❍
The new evaluation system accounts for the importance of content knowledge and content-specific pedagogy in evaluation	❍	❍	❍	❍	❍	❍
It is fair to be evaluated on my teaching by someone who is an expert on effective pedagogy even if they	❍	❍	❍	❍	❍	❍

are not familiar with my subject area.						
Effective teaching is generally the same across all content areas	❍	❍	❍	❍	❍	❍

Q23 In general, what kind of an effect do you think the new teacher evaluation system has had:

	Positive	Negative	No effect	Don't know
On your professional development	❍	❍	❍	❍
On collaboration with others	❍	❍	❍	❍
On your school	❍	❍	❍	❍

Q24 If you have any other comments or thoughts you would like to add which have not been captured by the previous questions, please write them below:

Appendix B: Year 2 Teacher Survey

Q2 Professional Background

Q3 Are you a teacher currently teaching students?

- ❍ Yes
- ❍ No

Q4 What is your gender?

- ❍ Male
- ❍ Female

Q5 What is the highest degree you have earned?

- ❍ No Degree
- ❍ Bachelor's degree
- ❍ Master's degree
- ❍ Doctorate
- ❍ Other

Q7 Which grades do you teach? Please select all that apply.

- ❑ Pre-K
- ❑ Kindergarten
- ❑ Grade 1
- ❑ Grade 2
- ❑ Grade 3
- ❑ Grade 4
- ❑ Grade 5
- ❑ Grade 6
- ❑ Grade 7
- ❑ Grade 8
- ❑ Grade 9
- ❑ Grade 10
- ❑ Grade 11
- ❑ Grade 12

Q8 Do you teach math or language arts in grades 4-8?

- ❍ Yes
- ❍ No

Q9 Do you teach one or more of the following core subjects? Please select all that apply.

- ❑ Math
- ❑ Language Arts (ELA)
- ❑ Science
- ❑ Social Studies

Q10 Do you teach any of the following:

- ❑ Special Education
- ❑ ELL
- ❑ Neither

Q11 What best describes your classroom instruction?

- ❍ Mostly self-contained or pull-out instruction
- ❍ Mostly inclusion instruction

Q12 How many years will you have been teaching at the end of the current school year ?

- ❍ Up to 3
- ❍ 4-8
- ❍ 9-15
- ❍ 16 or more

Q13 Do you have tenure at your present school?

- ❍ Yes
- ❍ No

Q14 Were you responsible for informing other teachers about the new teacher evaluation system?

- ❍ Yes
- ❍ No

Q15 How many hours of training or education have you personally received on the new teacher evaluation system:

- ❍ 0
- ❍ 1-3
- ❍ 4-5
- ❍ 6-8
- ❍ 9-16
- ❍ 17-24
- ❍ 25-32
- ❍ 33-40
- ❍ More than 40

Q16 What type of training did you receive on the teacher evaluation system and how much of it did you get?

	A lot	Some	Little	None
Video lectures	❍	❍	❍	❍
In-person lectures by the district	❍	❍	❍	❍
In-person lectures by outside experts	❍	❍	❍	❍
Opportunities for discussion with peers	❍	❍	❍	❍
Other (Please specify:)	❍	❍	❍	❍

Q17 Overall, how much training did you get on the teacher evaluation system?

- ❍ Too much
- ❍ Just enough
- ❍ Between enough and too little
- ❍ Too little
- ❍ Don't know

Q18 How many times have you been observed this school year?

- ❍ 0 times
- ❍ 1 time
- ❍ 2 times
- ❍ 3 times
- ❍ 4 times

Q19 How many dual observations (i.e. observed by two observers at the same time) have you had this school year?

❍ 0
❍ 1
❍ 2
❍ 3
❍ 4

Q20 How many of your observations have been conducted by an external observer (i.e. someone not from your school building)?

❍ 0
❍ 1
❍ 2
❍ 3
❍ 4

Q21 How many of your observations were unannounced?

❍ 0
❍ 1
❍ 2
❍ 3
❍ 4

Q22 How many of your observations were less than 30 minutes?

❍ 0
❍ 1
❍ 2
❍ 3
❍ 4

Q23 Have you had any involvement in developing student growth objectives (SGOs)?

- ❍ Yes
- ❍ No

Q24 Do you know how your summative evaluation will be calculated?

- ❍ Yes
- ❍ No

Q25 Please rate how much you agree with the following statements regarding the new teacher evaluation protocol (i.e., Danielson, Marzano, McREL, Stronge) in your school.

	Strongly agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree	Don't know / Not sure
a. The new teacher evaluation protocol can be used to generate accurate assessments.	❍	❍	❍	❍	❍	❍
b. The new teacher evaluation protocol can be used to clearly separate accomplished from unaccomplished teachers.	❍	❍	❍	❍	❍	❍
c. The criteria in the new teacher evaluation protocol can be used to appropriately assess teachers' performance.	❍	❍	❍	❍	❍	❍
d. The new teacher evaluation protocol can be used to assess what is important about teaching.	❍	❍	❍	❍	❍	❍

Q26 Please rate how much you agree with the following statements regarding the new teacher evaluation protocol (i.e., Danielson, Marzano, McREL, Stronge) in your school.

	Strongly agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree	Don't know / Not sure
a. The teacher evaluation protocol for assessing teachers is fair.	❍	❍	❍	❍	❍	❍
b. The teacher evaluation protocol takes into account the context of teaching.	❍	❍	❍	❍	❍	❍
c. The teacher evaluation protocol allows teachers to explain decisions and actions.	❍	❍	❍	❍	❍	❍
d. The teacher evaluation system provides a fair picture of my teaching.	❍	❍	❍	❍	❍	❍

Q27 How often do you receive the following as a result of the evaluation system?

	Very Frequently	Frequently	Sometimes	Rarely	Never
a. Feedback from student growth objectives (SGOs) or SGP	❍	❍	❍	❍	❍
b. Feedback from post-conferences	❍	❍	❍	❍	❍
c. Informal conversations with administrators	❍	❍	❍	❍	❍
d. Informal conversations with other teachers	❍	❍	❍	❍	❍
e. School-level feedback programs	❍	❍	❍	❍	❍
f. District-level feedback programs (e.g. workshops)	❍	❍	❍	❍	❍
g. Referral to outside sources of PD (e.g. articles, conferences, videos, etc.)	❍	❍	❍	❍	❍

Q28 Since the introduction of the new teacher observation system,

	Strongly agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree	Does not apply
a. ... the feedback I have received from observers has helped me to improve the learning environment in my classroom.	❍	❍	❍	❍	❍	❍
b. ... the feedback I have received from observers has helped me to improve the quality of my instruction.	❍	❍	❍	❍	❍	❍
c. ... I have changed my instructional methods to get higher ratings.	❍	❍	❍	❍	❍	❍

Q29 Please tell us how much you have changed your use of each of the practices listed below as aresult of the recent changes in teacher evaluation (both the new observation rubric and the newapproaches to assessing student growth) in your district.

	Greatly increased	Increased	No change	Decreased	Greatly decreased
a. Have students explain their conclusions, arguments, or solutions in writing.	❍	❍	❍	❍	❍
b. Teach strategies to score high on state or other standardized tests.	❍	❍	❍	❍	❍
c. Emphasize the importance of following procedures in solving mathematical problems or conducting scientific examinations.	❍	❍	❍	❍	❍
d. Have students show or explain a concept or idea in more than one way.	❍	❍	❍	❍	❍
e. Teach the regular curriculum using items like those on the state tests or locally developed	❍	❍	❍	❍	❍

final exams.					
f. Allow students to design their own project, experiment, or procedures for solving a problem.	❍	❍	❍	❍	❍
g. Have students use a rubric to grade each others' work.	❍	❍	❍	❍	❍
h. Motivate students to make their best efforts on the state tests or locally developed tests.	❍	❍	❍	❍	❍
i. Have students discuss or explain their ideas to each other in pairs or triplets.	❍	❍	❍	❍	❍

Q66 Please tell us how much you have changed your use of each of the practices listed below as aresult of the recent changes in teacher evaluation (both the new observation rubric and the newapproaches to assessing student growth) in your district.

	Greatly increased	Increased	No change	Decreased	Greatly decreased
a. Have students explain their conclusions, arguments, or solutions in writing.	❍	❍	❍	❍	❍
b. Teach strategies to score high on state or other standardized tests.	❍	❍	❍	❍	❍
c. Emphasize the importance of following procedures in solving mathematical problems or conducting scientific examinations.	❍	❍	❍	❍	❍
d. Have students show or explain a concept or idea in more than one way.	❍	❍	❍	❍	❍
e. Teach the regular curriculum using items like those on the state tests or locally developed	❍	❍	❍	❍	❍

final exams.					
f. Allow students to design their own project, experiment, or procedures for solving a problem.	❍	❍	❍	❍	❍
g. Have students use a rubric to grade each others' work.	❍	❍	❍	❍	❍
h. Motivate students to make their best efforts on the state tests or locally developed tests.	❍	❍	❍	❍	❍
i. Have students discuss or explain their ideas to each other in pairs or triplets.	❍	❍	❍	❍	❍

Q30 How important is each of the following factors in determining how you teach your students:

	Very important	Important	Neither Important nor Unimportant	Unimportant	Not at all important	Don't know
a. The new teacher observation process with its use of codified observation protocols	❍	❍	❍	❍	❍	❍
b. The use of SGPs and SGOs (student growth based on state tests or locally developed assessments) to evaluate you as a teacher	❍	❍	❍	❍	❍	❍
c. The development of an annual summative assessment of your teaching based on the teacher observation and student growth data	❍	❍	❍	❍	❍	❍
d. The practice of evaluating schools on the basis of	❍	❍	❍	❍	❍	❍

test scores and rating some as In Needs of Improvement						
e. The district curriculum	❍	❍	❍	❍	❍	❍
f. The textbooks and materials provided	❍	❍	❍	❍	❍	❍
g. The students in your class	❍	❍	❍	❍	❍	❍
h. Guidance from your principal	❍	❍	❍	❍	❍	❍
i. Guidance from your fellow teachers	❍	❍	❍	❍	❍	❍
j. Your personal beliefs and knowledge about good teaching	❍	❍	❍	❍	❍	❍

Q31 The introduction of student growth measures tied to teacher evaluation

	Strongly agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree	Does not apply
a. ...provides me with information that helps me to improve the learning environment in my classroom.	❍	❍	❍	❍	❍	❍
b. ...provides me with information that helps improve learning opportunities for my students.	❍	❍	❍	❍	❍	❍
c. ...has caused me to change my instructional methods to get higher ratings.	❍	❍	❍	❍	❍	❍

Q32 To what extent are these characteristics of the way student growth is assessed fair?

	Very fair	Fair	Neither fair nor unfair	Unfair	Very unfair
a. Some teachers get state-wide measures (SGPs) and others get locally developed measures (SGOs)	❍	❍	❍	❍	❍
b. Teachers have the opportunity to develop assessment items that reflect the content they teach	❍	❍	❍	❍	❍
c. Teachers have the opportunity to set cut points defining their own level of proficiency	❍	❍	❍	❍	❍
d. There is no agreed-upon way to assess proficiency in certain subjects (e.g.,art, music, physical education, etc.)	❍	❍	❍	❍	❍
e. Student characteristics like motivation or personal health may affect growth	❍	❍	❍	❍	❍

scores					

Q33 Please rate how much you agree with the following statements regarding the implementation of the new teacher evaluation pilot.

	Strongly agree	Agree	Neither agree nor disagree	Disagree	Strongly Disagree	Does not apply	Don't know
a. In my experience, the district's teacher evaluation protocol is well aligned with the district curriculum.	❍	❍	❍	❍	❍	❍	❍
b. The district's teacher evaluation protocol fits well with other school/district initiatives.	❍	❍	❍	❍	❍	❍	❍
c. The district's teacher evaluation protocol helps this district meet its accountability requirements under NCLB and other external mandates.	❍	❍	❍	❍	❍	❍	❍
d. The district's teacher evaluation protocol consumes	❍	❍	❍	❍	❍	❍	❍

resources that could be better spent on promoting key district improvement initiatives.							

Q34 Do you think that you have a chance of losing tenure under the new evaluation system?

- ❍ Very Likely
- ❍ Likely
- ❍ Neither likely nor unlikely
- ❍ Unlikely
- ❍ Very Unlikely

Q35 How do you think the new evaluation system will affect your chances of getting tenure?

- ❍ Greatly improve my chances
- ❍ Improve my chances
- ❍ Neither improve nor reduce my chances
- ❍ Reduce my chances
- ❍ Greatly reduce my chances
- ❍ Don't know

Q36 A major purpose of teacher evaluation in my district will be to provide information for teachers to improve their practice.

- ❍ Strongly agree
- ❍ Agree
- ❍ Neither Agree nor Disagree
- ❍ Disagree
- ❍ Strongly Disagree

Q37 A major purpose of teacher evaluation in my district will be to make tenure and promotion decisions.

- ❍ Strongly agree
- ❍ Agree
- ❍ Neither Agree nor Disagree
- ❍ Disagree
- ❍ Strongly Disagree

Q38 How important to you are the following factors when being evaluated under the new teacher evaluation rubric?

	Very important	Important	Neither important nor unimportant	Unimportant	Not at all important	Don't know
The person observing me knows my classroom well	❍	❍	❍	❍	❍	❍
The person observing me does not know me very well	❍	❍	❍	❍	❍	❍
The person observing me is from outside of my school building	❍	❍	❍	❍	❍	❍
The person observing me is from outside of my district	❍	❍	❍	❍	❍	❍

Q39 How important is to you that your observer has content knowledge in your content area?

- ❍ Very important
- ❍ Important
- ❍ Neither important nor unimportant
- ❍ Unimportant
- ❍ Not at all Important

Q40 Please rate your school broadly on the following dimensions:

	Excellent	Good	Inadequate	Poor
a. Technology	❍	❍	❍	❍
b. Personnel resources	❍	❍	❍	❍
c. Money for other purposes	❍	❍	❍	❍
d. Access to information about procedures and initiatives	❍	❍	❍	❍
e. Time to meet and prepare for class	❍	❍	❍	❍
f. Administrators' time to support teaching and learning	❍	❍	❍	❍

Q41 Please rate your level of agreement with the following statements.

	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree	Does not apply
a. I regularly participate in a scheduled planning period with other teachers.	❍	❍	❍	❍	❍	❍
b. I often meet with other teachers for a planning period.	❍	❍	❍	❍	❍	❍
c. A focused school vision for student learning is shared by most staff in the school.	❍	❍	❍	❍	❍	❍
d. There is a great deal of cooperative effort among staff members.	❍	❍	❍	❍	❍	❍
e. I make a conscious effort to coordinate the content of my courses with other teachers.	❍	❍	❍	❍	❍	❍

Q42 To what extent do you feel respected as a teacher...

	A great deal	Much	Somewhat	A little	Not at all	Don't know
a...by other teachers?	❍	❍	❍	❍	❍	❍
b...by your department chair?	❍	❍	❍	❍	❍	❍
c...by your principal (or equivalent)?	❍	❍	❍	❍	❍	❍
d...by your district office?	❍	❍	❍	❍	❍	❍
e...by students' parents?	❍	❍	❍	❍	❍	❍
f...by your students?	❍	❍	❍	❍	❍	❍
g...by your community?	❍	❍	❍	❍	❍	❍

Q43 Please rate your level of agreement with the following statements.

	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree	Don't know
a. It's OK in this school to discuss feelings, worries, and frustrations with the principal.	❍	❍	❍	❍	❍	❍
b. The principal looks out for the personal welfare of the faculty members.	❍	❍	❍	❍	❍	❍
c. I trust the principal at his or her word.	❍	❍	❍	❍	❍	❍
d. The principal has confidence in the expertise of the teachers.	❍	❍	❍	❍	❍	❍
e. The principal takes a personal interest in the professional development of the teachers.	❍	❍	❍	❍	❍	❍
f. I feel respected by the principal.	❍	❍	❍	❍	❍	❍

Q44 Please rate your level of agreement with the following statements.

	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree	Don't know
a. Teachers in this school trust each other.	❍	❍	❍	❍	❍	❍
b. It's OK in this school to discuss feelings, worries, and frustrations with other teachers.	❍	❍	❍	❍	❍	❍
c. Teachers respect other teachers who take the lead in school improvement efforts.	❍	❍	❍	❍	❍	❍
d. Teachers in this school really care about each other.	❍	❍	❍	❍	❍	❍
e.Most teachers in this school are cordial.	❍	❍	❍	❍	❍	❍

Q45 Please rate your level of agreement with the following statements.

	Strongly agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree	Don't know
a. My success or failure in teaching students is due primarily to factors beyond my control rather than to my own efforts and ability.	❍	❍	❍	❍	❍	❍
b. I sometimes feel it is a waste of time to attempt to do my best as a teacher.	❍	❍	❍	❍	❍	❍
c. I am certain I am making a difference in the lives of my students.	❍	❍	❍	❍	❍	❍
d. Many of the students I teach are not capable of learning the material I am supposed to teach them.	❍	❍	❍	❍	❍	❍
e. The	❍	❍	❍	❍	❍	❍

attitudes and habits my students bring to my class greatly reduce their chances for academic success.						

Q46 If you have any other comments or thoughts you would like to add which have not been captured by the previous questions, please write them below:

Appendix C: Year 1 Interview Guide

Teacher Focus Groups Interview Guide[7]

Questions About Observations

1. What have been your most important sources of knowledge about how to collect, analyze, and use teacher observation data? **(Probe: provider, state, written research, other consultants. Note: For teachers, ask "most important sources of knowledge about how the system works)**
2. In what ways is the teacher observation system useful (or not) for planning supervision, professional development, changes in the curriculum or other things?
3. **(FOR TEACHERS)** Thinking about the teacher observations that you have had this year, what made them better or worse than the observations you had last year?
4. How do you judge the expertise of the person who observes you. **(Probe: what competencies or kinds of knowledge are most critical for doing an accurate observation?)**
5. How well do you know the people who have observed you this year? **(Probe: how does your knowledge of the observer make you feel about the observations you experienced?)**
6. How is the teacher observation system facilitating or impeding collaboration among educators in this district?
7. What could be done to improve the usefulness of the teacher observation system?

Questions About Growth Scores

8. What progress has the district made in creating a system of growth scores for your students and classes?
9. STUDENT GROWTH SCORES (TESTED SUBJECTS)
   How are teacher growth scores facilitating or impeding collaboration among educators in this district?
10. TEST SCORES IN UNTESTED SUBJECTS
    How are measures of progress in untested areas facilitating or impeding collaboration among educators in this district?

[7] The interview guide shown here was used for teacher focus groups for the larger assessment of the EE4NJ pilot and not solely tailored to the analysis of this study.

Appendix D: Year 2 Focus Group Interview Guide

TEACHER FOCUS GROUP

Fall 2012

Note to interviewers: To conduct this focus group, you must have both the evaluation system and data storage system names prior to the focus group.

As you know, your district is participating in a teacher evaluation pilot this year. Next year, all districts in [STATE] will be required to formally evaluate teachers. We are trying to learn as much as we can from districts currently in the pilot so that we can help the state and all districts do a better job with teacher evaluation going forward.

1. I would like to start by having you go around the table and state the grade levels and subjects you teach.

2. How were teachers involved in deciding to pursue the teacher evaluation pilot program and in selecting ____ [Name of system] for use in this district?

3. Who would you say are the key leaders for the implementation of the teacher evaluation pilot in your school? In the district?

TEACHER OBSERVATION SYSTEM

4. To begin, I would like to discuss how you were trained on the new teacher observation system [e.g. Danielson, McREL etc.] (Facilitator: if specific points not addressed, raise again).

As we discuss this, please offer your insights on things such as:

a). Who did the training? How long was the training?

b). What has worked well about the training in the district on the observation system?

c). Was there something not covered in the training that you would like to have been covered?

d). What has NOT worked well about the training in the district on the observation system?

5. Next, I would like to discuss how you were trained on the [e.g.Teachscape, iObservation, etc.] system for recording teacher observation data.

As we discuss this, please offer your insights on things such as:

a). Who did the training? How long was the training?

b). Do you believe that after the training you understand how the system works and what is expected of you?

6. Next, I would like to discuss the actual observations you've had under the new system. It would be great if you can cover things like:

a). How many observations have you had thus far? By whom?

b). Who is doing the teacher observations?

c). Have you been observed by more than one observer? How often? Same or different lessons?

d). Have you experienced any unannounced visits? How did they differ for you from announced observations?

Once discussion above seems finished, probe further:

7. How accurate and reliable do you think observations are at this point? In what ways are observations accurate or not?

a). **Cohort 1**: in comparison with last year: better worse or the same?

8. Do you think that different observers in the district would be consistent in how they score the same observation?

9. Do you think that the observation scores that observers assign are consistent with how levels of teaching are defined in the x protocol? (Elaborate, that when observers assign a score of 4, the teaching that they observed matches the definition of a 4 in the protocol).

10. How fair do you think the observation system/pilot program is?

a). **Cohort 1**: in comparison with last year: better worse or the same?

11. How is feedback being provided to you?

a). How helpful has it been to you?

b). How has feedback been used for improvement? PROBE:

i. By yourself (have you used it?)

ii. By supervisors/administrators

12. When you think about the different tasks that go into teacher observation, which ones are the most time consuming for you?

13. How are you coping with the additional time requirements?

a). Is the additional work time, altered priorities, respondent's coping mechanisms worth it?

ABOUT OBSERVERS

14. When you think back on the observations that have been most accurate and insightful, what has the observer done that made the observations accurate and insightful? What individuals provided those observations?

a). Probe for any concerns about the accuracy of the observation scores in general, without getting into the scores you received?

15. What is the most important skill do you think a supervisor needs to have in order to complete quality teacher evaluation?

STUDENT DATA

16. Next, I would like to focus on student achievement goals (SGTs).

a). What are you doing in your school to develop SGTs? (Later in year) and collect student achievement data?

b). What SGTs have been identified? Who identified them? What considerations went into developing them? How will they be measured?

c). What do you need to know in order to choose appropriate SGTs, measure them, and analyze the data?

d). Has the process of developing SGTs been useful to you? In what ways?

RELATIONS

17. Has the implementation of the teacher evaluation pilot changed your relationship with your principals? Those who observe you?

18. How do you think observing and evaluating teachers differently will ultimately change student outcomes?

References

Amrein, A. L., & Berliner, D. C. (2002). High-stakes testing & student learning. *Education Policy Analysis Archives, 10*(18), 1-74. Retrieved September 8, 2014 from http://epaa.asu.edu/epaa/v10n18

Antonioni, D., & Park, H. (2001). The relationship between rater affect and three sources of 360-degree feedback ratings. *Journal of Management, 27*(4), 479-495.

Baker, B. D., & Oluwole, J. (2013). *Deconstructing disinformation on student growth percentiles & teacher evaluation in New Jersey.* Retrieved July 13, 2013, from http://njedpolicy.files.wordpress.com/2013/05/sgp_disinformation_bakeroluwole1.pdf

Baker, B. D., Oluwole, J., & Green, P. C. (2013). The legal consequences of mandating high stakes decisions based on low quality information: Teacher evaluation in the race-to-the-top era. *Educational Policy Analysis Archives, 21*(5), n5.

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., et al. (2010). Problems with the use of student test scores to evaluate teachers. EPI Briefing Paper# 278. *Economic Policy Institute*. Retrieved September 8, 2014 from http://files.eric.ed.gov/fulltext/ED516803.pdf

Baker, R., Walonoski, J., Heffernan, N., Roll, I., Corbett, A., & Koedinger, K. (2008). Why students engage in "gaming the system" behavior in interactive learning environments. *Journal of Interactive Learning Research, 19*(2), 185-224.

Bakkenes, I., Vermunt, J. D., & Wubbels, T. (2010). Teacher learning in the context of educational innovation: Learning activities and learning outcomes of experienced teachers. *Learning and Instruction, 20*(6), 533-548.

Ball, D. L., Thames, M. H., & Phelps, G. (2008). Content knowledge for teaching: What makes it special? *Journal of Teacher Education, 59*(5), 389-407.

Behind, N. C. L. (2002). Act of 2001, Pub. L. No. 107-110, § 115. *Stat, 1425*, 107-110.

Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., Pianata, R. C., & Qi, Y. (2012). An argument approach to observation protocol validity. *Educational Assessment, 17*(2-3), 62-87.

Betebenner, D. W. (2008). *A primer on student growth percentiles.* Dover, NH: National Center for the Improvement of Educational Assessment. Retrieved September 8, 2014 from https://www.gadoe.org/Curriculum-Instruction-and-Assessment/Assessment/Documents/Aprimeronstudentgrowthpercentiles.pdf

Betebenner, D. W. (2009). *Growth, standards, and accountability.* Dover, NH: National Center for the Improvement of Educational Assessment. Retrieved September 8, 2014 from http://www.nciea.org/publication_PDFs/growthandStandard_DB09.pdf

Bill and Melinda Gates Foundation. (2010). *Learning about teaching: Initial findings from the Measures of Teaching Project.* Retrieved September 8, 2014, from http://www.metproject.org/downloads/Preliminary_Findings-Research_Paper.pdf

Bill and Melinda Gates Foundation. (2010). *Working with teachers to develop fair and reliable measures of effective teaching.* Retrieved September 8, 2014, from http://www.metproject.org/downloads/met-framing-paper.pdf

Bill and Melinda Gates Foundation. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains.* Retrieved

October 1, 2012, from http://www.metproject.org/downloads/MET_Gathering_Feedback_Practioner_Brief.pdf

Bill and Melinda Gates Foundation. (2013). *Ensuring fair and reliable measures of effectve teaching: Culminating findings from the MET project's three-year study.* Retrieved September 8, 2014, from http://www.metproject.org/downloads/MET_Ensuring_Fair_and_Reliable_Measures_Practitioner_Brief.pdf

Brandt, C., Mathers, C., Oliva, M., Brown-Sims, M., & Hess, J. (2007). *Teacher evaluation practices in the Midwest Region (Issues & Answers Report, REL 2007-No. 030).* Washington, DC: U.S. Department of Education, Institute of Education Services, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Midwest. Retrieved from http://ies.ed.gov/ncee/edlabs.

Braun, H. I. (2005). Using student progress to evaluate teachers: A primer on value-added models. Policy Information Perspective. *Educational Testing Service*.

Brown, S., Seidelmann, A., & Zimmerman, G. (2006, July 1). *In the trenches: Three teachers' perspectives on moving beyond the Math Wars*. Retrieved June 26, 2014, from Mathematically Sane: http://mathematicallysane.com/in-the-trenches/

Bryk, A. S., & Schneider, B. (2003). Trust in schools: A core resource for school reform. *Educational Leadership, 60*(6), 40-44.

Caro, D. H. (2009). Socio-economic status and academic achievement trajectories from childhood to adolescence. *Canadian Journal of Education, 32*(3), 558-590.

Carpenter, T. P., Franke, M. L., Jacobs, V. R., Fennema, E., & Empson, S. B. (1998). A longitudinal study of invention and understanding in children's multidigit addition and subtraction. *Journal for Research in Mathematics Education, 29*(1), 3-20.

Cash, A. H., Hamre, B. K., Pianata, R. C., & Myers, S. S. (2012). Rater calibration when observational assessment occurs at large scale: Degree of calibration and characteristics of raters associated with calibration. *Early Childhood Research Quarterly, 27*(3), 529-542.

Coburn, C. E., & Turner, E. O. (2012). The practice of data use: An introduction. *American Journal of Education, 118*(2), 99-111.

Coffey, A., & Atkinson, P. (1996). *Making sense of qualitative data: Complementary research strategies.* Sage Publications, Inc.: Thousand Oaks, CA.

Conley, S., & Glasman, N. S. (2008). Fear, the School Organization, and Teacher Evaluation. *Educational Policy, 63*(1), 63-85.

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. Education Policy Action Series.* Providence, RI: Annenberg Institute for School Reform at Brown University (NJ1). Retrieved September 8, 2014 from http://annenberginstitute.org/pdf/valueaddedreport.pdf

Cronin, L., & Capie, W. (1986). *The influence of daily variation in teacher performance on the reliability and validity of assessment data.* University of Georgia. Retrieved September 8, 2014 from http://eric.ed.gov/?id=ED274704

Cruickshank, D. R., & Haefele, D. (2001, February). Good teachers, plural. *Educational Leadership, 58*(5), 26-30.

Cullen, J. B., & Reback, R. (2006). *Tinkering toward accolades: School gaming under a performance accountability system* (Vol. 14). Emerald Group Publishing Limited. Retrieved September 8, 2014 from http://www.nber.org/papers/w12286

Curtis, R. (2012). *Building it together: The design and implementation of Hillsborough County Public Schools' teacher evaluation system.* Washington, DC: The Aspen Institute. Retrieved September 8, 2014 from http://www.aspeninstitute.org/publications/building-it-together-design-implementation-hillsborough-county-public-schools-teacher

Danielson, C. (2007). *Enhancing professional practice: A framework for teaching.* Alexandria, Va: ASCD.

Danielson, C. (2011). Evaluations that help teachers learn. *Educational Leadership*, 68(4) 35-39.

Danielson, C. (2011). *The framework for teaching evaluation instrument: 2011 edition.* Princeton, NJ: The Danielson Goup.

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

Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E. H., & Rothstein, J. (2011). *Getting teacher evaluation right: A background paper for policy makers.* National Academy of Education (NJ1). REtrieved September 8, 2014 from http://tx.aft.org/files/gettingteacherevaluationright.pdf

Datnow, A., Park, V., & Kennedy-Lewis, B. (2012). High school teachers' use of data to inform instruction. *Journal of Education for Students Placed at Risk (JESPAR), 17*(4), 247-265.

Davis-Kean, P. E., & Sexton, H. R. (2009). Race differences in parental influences on child achievement: Multiple pathways to success. *Merrill-Palmer Quarterly (1982-)*, 285-318.

Dembrosky, J. W., Pane, J. F., Barney, H., & Christina, R. (2005). *Data driven decisionmaking in southwestern Pennsylvania school districts* . Santa Monica, CA: RAND.

Denner, P. R., Miller, T. L., Newsome, J. D., & Birdsong, J. R. (2002). Generalizability and validity of the use of a case analysis assessment to make visible the quality of teacher candidates. *Journal of Personnel Evaluation in Education, 16*(3), 153-174.

Donaldson, M. L., & Donaldson, Jr., G. A. (2012). Strengthening teacher evaluation: What district leaders can do. *Educational Leadership, 69*(8), 78-82.

Ehlert, M., Koedel, C., Parsons, E., & Podgursky, M. (2012). Selecting growth measures for school and teacher evaluations. *National Center for Analysis of Longitudinal Data in Education Research (CALDAR). Working Paper, 80*. Retrieved September 8, 2014 from http://economics.missouri.edu/working-papers/2012/WP1210_koedel.pdf

Firestone, W. A., Blitz, C. L., Gitomer, D. H., Gradinarova-Kirova, D., Shcherbakov, A., & Nordin, T. L. (2013). *External evaluation of the Excellent Educators for New Jersey Pilot evaluation: Year one final report.* Retrieved September 8, 2014, from http://www.state.nj.us/education/archive/EE4NJ/presources/RUGSE11-12.pdf

Firestone, W. A., Nordin, T. L., Shcherbakov, A., Kirova, D., & Blitz, C. L. (2014). *New Jersey's Pilot Teacher Evaluation Program: Year 2 final report.* Retrieved

September 8, 2014, from https://cesp.rutgers.edu/sites/default/files/RU-GSETEACHEREVALREPORT2014.pdf

Flores, M. A. (2010). Teacher Performance Appraisal in Portugal: The (Im)Possibilities of a Contested Model. *Mediterranean Journal of Educational Studies, 15*(1), 41-60.

Fullan, M. (2007). *The new meaning of educational change.* New York: Teachers College Press.

Gabriel, T. (2010, September 2). A Celebratory Road Trip by Education Secretary. *New York Times*, p. A24.

Ginsburg, M., Kamat, S., Raghu, R., & Weaver, J. (1995). Educators and politics: Interpretations, involvement, and implications. In M. B. Ginsburg (Ed.), *The politics of educators' work and lives* (pp. 3-54). New York: Garland Publishing.

Glasman, N. S., & Heck, R. (1987). Administrator engagement in evaluation for decisionmaking: The case of teacher assignment to classrooms. *Administrator's Notebook, 32*, 1-4.

Goldhaber, D., Walch, J., & Gabele, B. (2014). Does the model matter? Exploring the relationship between different student achievement-based teacher assessments. *Statistics and Public Policy, 1*(1), 28-39.

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

Goldstein, J., & Noguera, P. A. (2006, March). A thoughtful approach to teacher evaluation. *Educational Leadeship, 63*(6), 31-37.

Graham, M., Milanowski, A., & Miller, J. (2012). *Measuring and promoting inter-rater agreement of teacher and principal performance ratings.* Rockville, MD: Center for Educator Compensation Reform. Retrieved September 8, 2014 from http://cecr.ed.gov/pdfs/Inter_Rater.pdf

Haefele, D. L. (1993). Evaluating Teachers: A Call for Change. *Journal of Personnel Evaluation in Education, 7*(1), 21-31.

Hall, G. E., & Hord, S. M. (2006). *Implementing change: Patterns, principles, and potholes.* Boston, MA: Pearson/Allyn & Bacon.

Hall, G., & Hord, S. (1987). *Change in schools: Facilitating the process.* Albany, NY: SUNY Press.

Hallinger, P. (2003). Leading educational change: Reflections on the practice of instructional and transformational leadership. *Cambridge Journal of Education, 33*(3), 329-352.

Hamilton, L., Halverson, R., Jackson, S. S., Mandinach, E., Supovitz, J. A., & Wayman, J. C. (2009). Using student achievement data to support instructional decision making. IES Practice Guide (NCEE 2009-4067). *National Center for Education Evaluarion and Regional Assistance, Institute of Education Services, U.S. Department of Education*. Retrieved September 8, 2014 from http://ies.ed.gov/ncee/wwc/PracticeGuide.aspx?sid=12

Hanushek, E. A. (1997). Assessing the effects of school resources on student performance: An update. *Educational Evaluation and Policy Analysis, 19*(2), 141-164.

Hanushek, E. A., & Rivkin, S. G. (2010). Using value-added measures of teacher quality. Brief 9. *National Center for Analysis of Longitudinal Data in Education*

Research. Retrieved September 8, 2014 from http://www.urban.org/uploadedpdf/1001371-teacher-quality.pdf

Hargreaves, A. (2005). Educational change takes ages: Life, career and generational factors in teachers' emotional responses to educational change. *Teaching and Teacher Education, 21*(8), 967-983.

Harris, D. N., & Sass, T. R. (2009). *What makes for a good teacher and who can tell?* Retrieved September 8, 2014, from http://www.urban.org/uploadedpdf/1001431-what-makes-for-a-good-teacher.pdf

Heck, R., & Hallinger, P. (2009). Assessing the contribution of principal and teacher leadership to school improvement. *American Educational Research Journal, 46*(3), 659-689.

Hill, H. C., Kapitula, L., & Umland, K. (2011). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal, 48*(3), 794-831.

Hoy, W. K., & Sweetland, S. R. (2001). Designing better schools: The meaning and measure of enabling school structures. *Educational Administration Quarterly, 37*(3), 296-321.

Huberman, M. (1983). Recipes for busy kitchens. *Science Communication, 4*(4), 478-510.

Ikemoto, G. S., & Marsh, J. A. (2007). Cutting through the "Data-Driven" mantra: Different conceptions of Data-Driven Decision Making. *Yearbook of the National Society for the Study of Education, 106*(1), 105-131.

Jennings, J. (2012). The effects of accountability system design on teachers' use of test score data. *Teachers College Record, 114*(11), 1-23.

Knapp, M. S., Copland, M. A., & Talbert, J. E. (2003). *Leading for Learning: Reflective tools for school and district leaders.* Seattle, WA: Center for the Study of Teaching and Policy, University of Washington. Retrieved September 8, 2014 from https://depts.washington.edu/ctpmail/PDFs/LforLSummary-02-03.pdf

Kohn, A. (2000). *The schools our children deserve: Moving beyond traditional classrooms and "tougher standards"*. Houghton Mifflin Harcourt.

Koretz, D. (2008). *Measuring up: what standardized testing really tells us.* Cambridge, MA: Harvard University Press.

Kulm, G., & Stuessy, C. (1991). Assessment in science and mathematics education reform. In G. Kulm, & S. Malcom (Eds.), *Science assessment in the service of reform* (pp. 71-87). Washington DC: American Association for the Advancement of Science.

La Paro, K. M., Pianata, R. C., & Stuhlman, M. (2004). The classroom assessment scoring system: Findings from the prekindergarten year. *The Elementary School Journal, 104*(5), 409-426.

Leithwood, K. (1992). The principal's role in teacher development. In M. Fullan, & A. Hargreaves (Eds.), *Teacher development and educational change* (pp. 86-103). Lewes: Falmer Press.

Leithwood, K., & Jantzi, D. (1999). Transformational school leadership effects: A replication. *School Effectiveness and School Improvement, 10*(4), 451-479.

Leithwood, K., & Jantzi, D. (2000). The effects of transformational leadership on organizational conditions and student engagement with school. *Journal of Educational Administration, 38*(2), 112-129.

Leithwood, K., & Jantzi, D. (2005). A review of transformational school leader research: 1996-2005. *Leadership and policy in schools, 4*(3), 177-199.

Lieberman, A., & Mace, D. H. (2008). Teacher learning: The key to educational reform. *Journal of Teacher Education, 59*(3), 226-234.

Lipsky, M. (2010). *Street-level democracy: Dilemmas of the individual in public services. 30th anniversary edition.* New York: Russel Sage Foundation.

Little, O., Goe, L., & Bell, C. (2009). *A practical guide to evaluating teacher effectiveness.* Washington, DC: National Comprehensive Center for Teacher Quality. Retrieved September 8, 2014 from http://files.eric.ed.gov/fulltext/ED543776.pdf

Lizzio, A., Wilson, K., & Que, J. (2009). Relationship dimensions in the professional supervision of psychology graduates: Supervisee perceptions of process and outcome. *Studies in Continuing Education, 31*(2), 127-140.

Louis, K. S. (2007). Trust and improvement in schools. *Journal of Educational Change, 8*(1), 1-24.

Mandinach, E. B., & Honey, M. (2008). *Data driven school improvement: Linking data and learning.* New York, NY: Teachers College Press.

Mandinach, E., Honey, M., Light, D., & Brunner, C. (2008). A conceptual framework for data-driven decision-making. In E. Mandinach, & M. Honey (Eds.), *Data-driven school improvement: Linking data and learning* (pp. 13-31). New York, NY: Teachers College Press.

Marsh, J. A. (2012). Interventions promoting educators' use of data: Research insights and gaps. *Teachers College Record, 114*(11), 1-48.

Marsh, J. A., Pane, J. F., & Hamilton, L. S. (2006). *Making sense of data-driven decision making in education.* RAND Corporation. Retrieved September 8, 2014 from http://www.rand.org/content/dam/rand/pubs/occasional_papers/2006/RAND_OP170.pdf

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

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

Mason, S. (2002). *Turning data into knowledge: Lessons from six Milwaukee public schools.* Madison, WI: Wisconsin Center for Education Research. Retrieved September 8, 2014 from http://www.wcer.wisc.edu/archive/mps/AERA2002/Mason%20AERA%202002%20QSP%20Symposium%20Paper.pdf

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

Milanowski, A., & Kimball, S. M. (2004). *The framework-based performance assessment systems in Cincinnati and Washoe.* Retrieved September 8, 2014, from http://www.cpre-wisconsin.org/papers/CinciWashoe_TE.pdf

Miles, M. B., & Huberman, A. M. (1994). *Qualitative Data Analysis, Second Edition.* Thousand Oaks, CA: Sage.

Montgomery, D. (2013). *Helping teachers develop through classroom observation, second edition.* Routledge.

National Comission on Excellence in Education, The. (1983). *A nation at risk: The imperative for educational reform.* Washington, DC: United States Department of Education.

Nevo, D. (1994). How can teachers benefit from teacher evaluation. *Journal of Personnel Evaluation in Education, 8*(2), 109-117.

New Jersey Department of Education. (2013, May 7). *AchieveNJ: Educator Evaluation and Support in New Jersey.* Retrieved June 19, 2013, from http://www.state.nj.us/education/AchieveNJ/intro/guide.pdf

New Jersey Educator Effectiveness Task Force. (2011). *Notice of grant opportunity. Excellent Educators for New Jersey (EE4NJ) pilot program.* Retrieved September 8, 2014, from http://www.state.nj.us/education/grants/docs/11-CO01-S01.doc

Nolan Jr., J., & Hoover, L. A. (2004). *Teacher supervision and evaluation.* Hoboken, NJ: John Wiley and Sons.

Norcross, J. C., & Haglin, R. P. (1997). Integrative approaches to psychotherapy supervision. In C. E. Watkins, Jr. (Ed.), *Handbook of psychotherapy supervision* (pp. 508-524). New York: Wiley.

Oliva, M., Mathers, C., & Laine, S. (2009, March). Effective evaluation. *Principal Leadership, 9*(7), pp. 16-21.

Olson, D. R. (2004). The triumph of hope over experience in the search for "what works": A response to Slavin. *Educational Researcher, 33*(1), 24-26.

Park, V., & Datnow, A. (2009). Co-constructing distributed leadership: District and school connections in data-driven decision making. *School Leadership and Management, 29*(5), 477-494.

Patton, M. Q. (1990). *Qualitative evaluation and research methods.* Newbury Park: Sage.

Perez-Pena, R. (2011, January 11). *Christie calls for overhaul of education*. Retrieved February 11, 2013, from New York Times: http://www.nytimes.com/2011/01/12/nyregion/12jersey.html?_r=0

Pesek, D. D., & Kirshner, D. (2000). Interference of instrumental instruction in subsequent relational learning. *Journal for Research in Mathematics Education, 31*(5), 524-540.

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

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

Popham, W. J. (1988). The dysfunctional marriage of formative and summative teacher evaluation. *Journal of Personnel Evaluation in Education, 1*(3), 269-273.

Postman, N. (2011). *The end of education: Redefining the value of school.* Random House LLC.

Prestine, N. A., & Nelson, B. S. (2005). How can educational leaders support and promote teaching and learning? New conceptions of learning and leading in schools. In W. A. Firestone, & C. Riehl (Eds.), *A New Agenda for Research in Educational Leadership* (pp. 46-60). New York: Teachers College Press.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2001). *Teachers, schools, and academic achievement.* Amherst, MA: Amherst College.

Robbins, T. L., & DeNisi, A. S. (1994). A closer look at interpersonal affect as a distinct influence on cognitive processing in performance evaluations. *Journal of Applied Psychology, 79*(3), 341-353.

Rowan, B., Correnti, R., & Miller, R. J. (2006). What large scale survey research tells us about teacher effects on student achievement: Insights from the Prospects study of elementary schools. *Teachers College Record, 104*(8), 1525-1567.

Rubin, K. H., Bukowski, W., & Parker, J. G. (2006). Peer interactions, relationships, and groups. In N. Eisenberg (Ed.), *Handbook of child psychology: Social, emotional, and personality development* (Vol. 3, pp. 571-645). New York: Wiley.

Sartain, L., Stoelinga, S. R., & Krone, E. (2010). *Rethinking teacher evaluation: Findings from the first year of the Excellence in Teaching Project in Chicago Public Schools. Policy Brief.* Chicago: Consortium on Chicago School Research. Retrieved September 8, 2014 from http://www.isbe.net/peac/pdf/teacher_eval_brief.pdf

Sawchuck, S. (2013). *High ratings for teachers still seen.* Retrieved September 8, 2014, from Education Week: http://www.edweek.org/ew/articles/2013/02/06/20evaluate_ep.h32.html

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher, 15*(2), 4-14.

Slavin, R. E. (2002). Evidence-based education policies: Transforming practice and research. *Educational Researcher, 31*(7), 15-21.

Stein, M. K., & Nelson, B. S. (2003). Leadership content knowledge. *Educational Evaluation and Policy Analysis, 25*(4), 423-448.

Stigler, J. W., & Hiebert, J. (2009). *The teaching gap: Best ideas from the world's teachers for improving education in the classroom.* Simon and Schuster.

Stronge, J. H. (2012). *Stronge teacher effectiveness performance evaluation system.* Williamsburg, VA: Stronge & Associates Educational Consultants, LLC.

Stronge, J. H., & Tucker, P. D. (1999). The politics of teacher evaluation: A case study of new system design and implementation. *Journal of Personnel Evaluation in Education, 13*(4), 339-359.

The Danielson Group. (2011). *The framework for teaching*. Retrieved January 15, 2013, from The Danielson Group: http://www.danielsongroup.org/article.aspx?page=frameworkforteaching

Toch, T., & Rothman, R. (2008). *Rush to judgment: Teacher evaluation in public education. Education Sector Reports.* Washington, DC: Education Sector. Retrieved September 8, 2014 from http://www.educationsector.org/publications/rush-judgment-teacher-evaluation-public-education

Tschannen-Moran, M. (2004). *Trust matters: Leadership for successful schools.* San Francisco, CA: Jossey-Bass.

Tyack, D., & Cuban, L. (1995). *Tinkering toward utopia: A century of public school reform.* Cambridge, MA: Harvard University Press.

Tynjälä, P. (2008). Perspectives into learning at the workplace. *Educational Research Review, 3*(2), 130-154.

United States Department of Education Office for Planning Evaluation and Policy Development. (2010, March). *ESEA: A blueprint for reform: The reauthorization of the Elementary and Secondary Education Act*. Retrieved October 12, 2012, from ed.gov: http://www2.ed.gov/policy/elsec/leg/blueprint/blueprint.pdf

Van Driel, J. H., & Berry, A. (2010). Pedagogical content knowledge. In P. L. Peterson, E. L. Baker, & B. McGraw (Eds.), *International Encyclopedia of Education Third Edition* (pp. 656-661). Elsevier Science Limited.

Wahlstrom, K. L., & Louis, K. S. (2008). How teachers experience principal leadership: The roles of professional community, trust, efficacy, and shared responsibility. *Educational Administration Quarterly, 44*(4), 458-395.

Watkins, C. E. (1997). Defining psychotherapy supervision and understanding supervisor functioning. In C. Watkins, Jr. (Ed.), *Handbook of psychotherapy supervision* (pp. 3-10). New York: Wiley.

Wayne, A. J., & Youngs, P. (2003). Teacher characteristics and student achievement gains: A review. *Review of Educational Research, 73*(1), 89-122.

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness.* New York, NY: The New Teacher Project. Retrieved September 8, 2014 from http://tntp.org/assets/documents/TheWidgetEffect_2nd_ed.pdf

Williams, J. (2009). *McREL's Teacher evaluation framework.* Denver, CO: McREL.

Wilson, K. J., Charker, J., Lizzio, A., Halford, K., & Kimlin, S. (2005). Assessing how much people work at their relationships: The Behavioural Self-Regulation for Effective Relationships Scale. *Journal of Family Psychology, 19*, 385-393.

Wohlstetter, P., Datnow, A., & Park, V. (2008). Creating a system for data-driven decision-making: Applying the principal-agent framework. *School Effectiveness and School Improvement, 19*(3), 239-259.

Worthen, V., & McNeil, B. W. (1996). A phenomenological investigation of 'good' supervision events. *Journal of Counseling Psychology, 43*(1), 25-34.

Yin, R. (2008). *Case study research: Design and methods.* Thousand Oaks, CA: Sage.

Yuan, K., Le, V.-N., McCaffrey, D. F., Marsh, J. A., Hamilton, L. S., Stecher, B. M., et al. (2013). Incentive pay programs do not affect teacher motivation or reported practices: Results from three randomized studies. *Educational Evaluation and Policy Analysis, 35*(1), 3-22.

Zeichner, K. M., & Liston, D. P. (2013). *Reflective teaching: An introduction.* Routledge.