Spring 5-12-2018

Teacher Performance Evaluation and Professional Growth in the Era of "Educator Effectiveness" in Maine

Jonathan E. Doty
University of Maine, jon.e.doty@gmail.com

TEACHER PERFORMANCE EVALUATION AND PROFESSIONAL GROWTH IN THE ERA OF "EDUCATOR EFFECTIVENESS" IN MAINE

By

Jonathan Edwin Doty

B.S. University of Maine, 2000

M.Ed. University of Maine, 2004

C.A.S. University of Maine, 2006

A DISSERTATION

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Education

(in Educational Leadership)

The Graduate School

The University of Maine

May 2018

Advisory Committee:

Ian Mette, Assistant Professor of Educational Leadership, Advisor
Richard Ackerman, Professor of Educational Leadership
Janet Fairman, Associate Professor of Education
Tammy Mills, Assistant Professor of Curriculum, Assessment, and Instruction
Rebecca Schwartz-Mette, Assistant Professor of Psychology

TEACHER PERFORMANCE EVALUATION AND PROFESSIONAL GROWTH IN THE ERA OF "EDUCATOR EFFECTIVENESS" IN MAINE

By Jonathan Edwin Doty

Dissertation Advisor: Dr. Ian Mette

An Abstract of the Dissertation Presented
in Partial Fulfillment of the Requirements for the
Degree of Doctor of Education
(in Educational Leadership)
May 2018

Maine is one of many states that undertook reform to teacher supervision and evaluation in the wake of public attention (e.g., *Waiting for Superman*) and federal pressure (e.g., NCLB Flexibility Waivers). The Maine Legislature passed An Act to Ensure Effective Teaching and School Leadership (2012), shifting from local discretion to greater state influence on the functions of formative supervision and summative evaluation. As school districts created systems to meet the state's mandates, they combined growth and employment functions and navigated persistent challenges described in the literature on evaluation and supervision.

The purpose of this study was to examine perspectives from the field as to major local changes in teacher performance evaluation (PE) and professional growth (PG), the ways in which local PE & PG systems were or were not beginning to improve teacher effectiveness, and perceptions of factors contributing to or providing barriers to this improvement. This mixed-method, multi-site case study captured eight school districts as they piloted or implemented systems; the sites were purposefully selected to yield a rural and a non-rural site for each of the professional practice models frequently chosen in Maine. Teachers, evaluators, and supervisors were interviewed (20 total); 302 practitioners in the same roles contributed survey data.

Data were analyzed through multi-cycle coding (Saldaña, 2016), descriptive statistics, and basic inferential statistics. Major changes were underway, including implementation of new and more detailed professional standards, rubrics, and processes for supervision and evaluation. Sites were striving to put professional growth in the forefront and were perceiving positive gains with the detailed standards and cultural efforts, but some intentions such as increasing formative feedback to teachers were not yet realized. New resources were rare and the implementation of sites' aspirations exposed the scarcity of time for all involved, especially for evaluators (e.g., Principals). Participants largely found the student growth data evaluation mandate unhelpful. Overall in this piloting and early implementation stage participants were not yet seeing the intended increase in effectiveness, but promising practices emerged along with rural differences and the need to address the scarcity of time for evaluative accuracy and formative growth.

DEDICATION

To educators everywhere who work hard day in and day out for their students.

ACKNOWLEDGMENTS

First and foremost, I thank my family for their inspiration, patience, and support: my parents Ruth and Richard, my brother Matt and sister-in-law Heather, and especially my wife Kristen (who deserves a full acknowledgments page). Educational aspirations, expectations, and support have been priorities in our family for generations; Kristen and I seek to pass that on to our children, Lucas and Lauren.

Beyond family, mentors and friends such as Craig and MaryBeth Mucher, Josh Puhlick, and Gail Garthwait have contributed to my journey at points along the road.

I have been very fortunate to have grown along with Old Town School Department and RSU #34 for my entire career, where promotion of and support for lifelong learning have been constants. I have been surrounded by tremendous and inspiring educators. I especially wish to recognize John Keane for engaging me in continued study and many great debates of education, Judy Pusey for challenging me to earn the doctorate, and David Walker and Jeanna Tuell for supporting and encouraging me throughout that journey.

I am grateful to my advisor, Ian Mette, and the faculty who have provided guidance, challenge, critique, and support during my coursework and the design and implementation of this study: Richard Ackerman, Rebecca Buchanan, Gordon Donaldson, Janet Fairman, Susan Gardner, Diane Hoff, Paul Knowles, Sally Mackenzie, George Marnik, Tammy Mills, and Rebecca Schwartz-Mette.

Finally, I greatly appreciate the educators who participated in my study, taking time and energy from their busy days and lives to contribute their experiences and perspectives.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

THE PROBLEM

Teacher evaluation and supervision reform is underway nationwide, prompted in part by pressure from the federal No Child Left Behind (NCLB) flexibility waivers and the Race to the Top program (Darling-Hammond, 2013). Maine is a good example of this reform, beginning a shift in 2012 away from a long history of local (school district) control of teacher[1] supervision and evaluation. Maine's legislated requirements for Performance Evaluation and Professional Growth Systems (PE & PG Systems) combined formative and summative functions (i.e., teacher supervision and teacher evaluation), with seemingly a much more intensive focus on summative functions (State of Maine, 2015). Maine school districts were then obligated to revise old systems or create new systems with numerous required processes and outcomes, along with the challenge of addressing well-known tensions of role and authority when supervising and evaluating teachers (Oliva & Pawlas, 2004). The legislated changes did not include significant financial or personnel resources for local school districts to deploy in developing and implementing complex policy (a grant of $4,600.00 was made available to each district).

A climate of continual change and unpredictability influenced implementation, with changes to Maine statute and regulation in 2012, 2014, 2015, and 2016 (see timeline in Table 1.1 below). Further influencing unpredictability, the federal passage of the *Every Student Succeeds Act* (ESSA), which replaced NCLB, NCLB Flexibility Waivers, and the Race to the Top program, removed much of the federal pressure for inclusion of certain elements in state teacher

[1] These Maine laws and regulations also address principal Performance Evaluation and Professional Growth Systems; this study focuses only on the teacher components.

Table 1.1. *Major Events Leading to PE & PG Implementation*

Date	Level	Action
2002 (Jan.)	Federal	*No Child Left Behind Act of 2001* Becomes Law
2009 (Nov.)	Federal	*Race to the Top* Fund Notices Published
2011 (Sept.)	Federal	*No Child Left Behind* Flexibility Waivers Announced
2012 (April)	Maine	LD 1858 Becomes Law, "*An Act to Ensure Effective Teaching and School Leadership*"
2014 (May)	Maine	LD 1747 Becomes Law, "*Resolve, Regarding Legislative Review of Chapter 180: Performance Evaluation and Professional Growth Systems.*" Chapter 180 Takes Effect June 20th.
2015 (March)	Maine	LD 692 Becomes Law, "*An Act Regarding Educator Effectiveness.*" Chapter 180 Amended
2015 (April)	Maine	LD 38 Becomes Law, "*An Act to Allow Sufficient Time for Implementation of the Performance Evaluation and Professional Growth System for Educators*"
2015 (May)	Maine	LD 461 Becomes Law, "*An Act to Change the Notification Deadline for the Nonrenewal of a Teacher's Contract*"
2015 (July)	Local	Pilot Plans Due, Stakeholder Groups Must Reach Consensus Regarding Most Issues or Use State Defaults
2015 (Sept.)	Local	PE & PG System Pilot for One or More Schools
2015 (Dec.)	Federal	*Every Student Succeeds Act* Becomes Law
2016 (March)	Maine	LD 1459 Becomes Law, "*An Act to Clarify the Use of Student Data from the Statewide Assessment Test*"
2016 (Sept.)	Local	PE & PG System Pilot for All Schools and Applicable Staff
2016 - 2017	**Local**	**Data Collection for This Study (Nov. 2016 – June 2017)**
2017 (Sept.)	Local	PE & PG System Implementation

Note. At local pilot stages districts had the option to implement if ready.

evaluation and growth systems (Posey, 2016). This raised the possibility that Maine may make further revisions to statute and regulation.

Statement of the Problem

The stated purpose of Maine's changes to teacher supervision and evaluation systems is "to improve educator effectiveness by clearly setting forth expectations for professional practice and student learning and growth, and providing actionable feedback and support to help educators meet those expectations" (2015, Section 1). On their face the Maine governmental actions (under federal pressure) requiring certain elements in teacher performance evaluation and professional growth appear rational, with a focus on standards and improvement to reach those standards. The majority of elements in Maine's law and regulation (described in detail later in this chapter) were non-controversial to the field and the public; the required use of student growth data in evaluation drew the most debate.

However, as Hallinger, Heck, and Murphy (2014) concluded in their national research review described in Chapter 2, "the policy logic supporting [such] reform remains considerably stronger than the empirical evidence" (p. 5). In the piloting and early rollout of Maine's new Performance Evaluation and Professional Growth (PE & PG) systems, an opportunity existed to conduct research that addressed a problem affecting the field and the literature: Little research existed nationwide about this policy shift that impacted many states, particularly in a context like Maine's, where local districts retained some latitude in key elements of the system (e.g., professional practice model, weighting of system components). Further, what research had been conducted was almost entirely from the perspectives of evaluators[2] of teachers (e.g., principals)

[2] Throughout this study, the term *evaluator* is used to mean one responsible for directly evaluating teacher performance for employment purposes, e.g., a principal or director.

rather than of teachers or non-evaluative supervisors of teachers (e.g., Goldring et al., 2015; Mason & Porter, 2015). There was a need, then, to provide further study that contributed to the literature and informed practitioners and policymakers as they undertook implementation and considered refinement of these complex and important systems. In the remainder of this section I provide more detail to support this need.

First, on both the national and state levels, research regarding systems of teacher evaluation and teacher supervision in the context of recent federal pressures is limited. The 2009 *Race to the Top* program and the 2011 *NCLB Flexibility Waivers* moved some 24 states to take action on teacher evaluation systems (Kraft & Gilmour, 2017). The approaches varied from state to state. For example, Ohio has a common set of teacher standards and dictates the evaluative weight of student growth data (Kowalski & Dolph, 2015), while Maine leaves each to local discretion within some parameters. Much has been written in the theoretical literature about what educators and scholars think makes for effective supervision (Glickman, Gordon, & Ross-Gordon, 2013; Zepeda, 2012), evaluation (Darling-Hammond, 2013; Marzano & Toth, 2013; Platt, Tripp, Ogden, & Fraser, 2000), and the merging of these two functions (Darling-Hammond, 2013; Marshall, 2013; Marzano & Toth, 2013; Mette et al., 2017) in the best interest of students. However, the research base addressing implementation of these state mandates that also include student growth data in evaluation is slowly emerging, and has few contributions that include perspectives from roles other than evaluators of teachers (for one example see Ritter & Barnett's 2016 study in Tennessee).

In Maine, the Maine Educational Policy Research Institute's (MEPRI) studies were the only published evidence found as part of the literature review for this study. These MEPRI studies included a statewide survey (n = 76) and qualitative case studies (n = 7) yielding the

perspectives of superintendents or designees (Mason & Porter, 2014; Mason & Tu, 2015), the perspectives of school or district administrators from four school districts in the piloting year (Mette & Fairman, 2016), and the perspectives of school or district administrators from six school districts in the first year of implementation (Fairman & Mette, 2017). Thus, similar to the national literature, research on the effects of Maine's mandates was in need of additional contributions, especially to include non-evaluators' perspectives.

On the ground in Maine, school districts were nearly-simultaneously undertaking substantial change in the wake of several dynamic years of policy reforms (outlined in Table 1.1 above). At this point in time it was too early to study whether teacher effectiveness had actually increased under the new Maine systems. However, after several years of statewide reforms, and at the conclusion of the first state-mandated pilot year, it was possible to begin to address the gap in knowledge about the local changes, whether summative evaluation or formative growth were in the forefront of local efforts, and what factors were influencing progress toward the desired outcome of improved educator effectiveness.

What was already known about Maine implementation of PE & PG systems indicated a need for continued study. As part of the limited research in Maine regarding school district responses to Chapter 180, Mette and Fairman (2016) and Fairman and Mette (2017) conducted interviews with school and district leaders about their experiences piloting and implementing components of PE & PG systems. They found local practitioners shared some successes, such as "increased clarity in professional practice standards" (2016, p. i) and success in conducting small-scale pilots before expanding the scope of new systems. The same local leaders shared challenges including "difficulty and uncertainty about measuring student growth," a "substantial time commitment for district and school administrators," and state-level "ongoing changes to

PE/PG and inconsistencies" (2016, p. ii). These findings supported my personal experiences and anecdotes shared with me by school and district leaders that factors impacting improvement of effectiveness likely included, but were not limited to, the total weight of regulatory detail, the unpredictability as the law and regulation progressed through several revisions (and as future revisions were anticipated), tension in combining (rather than separating) the different functions of supervision and evaluation, additional time-consuming requirements placed on busy school staff, multiple external parallel changes in state assessment, and the challenge of revamping decades-old systems and cultures.

Across the nation and across Maine, a new era of legislatively combined supervision and evaluation was underway. Little was known about the local design and implementation of the mandated systems, especially regarding the perspectives of non-evaluators, creating a need in the literature. For practitioners, additional research would add ability to learn from others' implementations and apply lessons learned. For policymakers, additional research would contribute data and perspectives that could inform staying the course, adjustment of policy, or efforts to address factors arising as barriers to the policy goals.

Purpose

Thus, there was a need to understand more about the initial local (school district) implementation of this law and regulation. Specifically, the purpose of this study was to examine perspectives from the field regarding major local changes in teacher performance evaluation and teacher professional growth, perceptions of the ways in which local PE & PG systems were or were not beginning to improve teacher effectiveness, and perceptions of factors contributing to or providing barriers to this improvement. By examining the early implementation of PE & PG systems the researcher sought to better understand what was

happening at the local level through the perceptions of practitioners (teachers, supervisors, and evaluators). As a policy combining supervision and evaluation was implemented into practice, this increased understanding additionally allows the researcher to inform practitioners and researchers about supervision and evaluation functions critical to teachers, leaders, and students.

Research Questions

The study centered on three primary research questions examining the early local (school district) piloting and implementation of the teacher components of Maine's 2012 Educator Effectiveness law, 2014 Performance Evaluation and Professional Growth Systems regulation, and 2015–2016 revisions to law and regulation:

1. What did practitioners (teachers, supervisors, and evaluators) perceive as major changes in teacher performance evaluation and teacher professional growth in their school or district?
2. What factors did practitioners perceive as contributing to improved teacher effectiveness via teacher performance evaluation and/or teacher professional growth in their school or district?
3. What factors did practitioners perceive as barriers to improved teacher effectiveness via teacher performance evaluation and/or teacher professional growth in their school or district?

Supervision and Evaluation Defined

Clear definition and careful review of the literature regarding teacher evaluation and supervision for teacher growth are necessary in order to understand the reform efforts local school districts are making in response to Maine's mandates, the preliminary outcomes of those efforts, and the perceptions of teachers and administrators who are affected by those outcomes. I

present here a brief overview of these functions, and expand in detail in Chapter 2. The meaning of evaluate is "to judge the value or condition of (someone or something) in a careful and thoughtful way" (Merriam-Webster, 2016). With regards to teachers, performance evaluation (i.e., summative evaluation) is primarily an administrative function, to determine whether the teacher meets minimum standards and in some situations to determine positive or negative employment actions (Glickman et al., 2013). Teacher evaluation, as detailed in the review of the literature, is complex and problematic.

Interestingly enough, the meaning of supervise is "to coordinate and direct the activities of" (Merriam-Webster, 2016). With regards to teachers, this meaning is typically expanded to include support for professional growth (Glickman et al., 2013). As detailed in the review of the literature, teacher supervision is complex and situational, relying on a combination of knowledge, technical skill, and interpersonal skill (e.g., Zepeda, 2012; Glickman et al., 2013). The terms are often times misused in the field, as evaluation and supervision are inherently different and lead to different results, mainly that evaluation is a function to determine employment consequences and supervision is a function to provide direction and support for growth. Under Maine's mandate, school districts are charged with developing and implementing criteria for both functions in a combined PE & PG system.

The Context of the Problem

In this section I describe the context in which the problem emerged, and in which I conducted this study. Prior to 2012 Maine statute and regulation left much discretion to the local school district with regards to teacher supervision and evaluation. In this regard Maine reinforced a tradition of local control ingrained in many local communities (Jimerson, 2005). The 125th Maine Legislature passed An Act to Ensure Effective Teaching and School Leadership

(2012), beginning an era of greater state influence on both functions. The Legislature further refined this influence with multiple revisions to regulation and statute (laid out in Table 1.1 above).

With those acts the Maine Legislature shifted from what Hazi and Ricinski (2009) defined as the lowest level of state control in teacher evaluation ("least prescriptive…local discretion") to a much higher level: "definitional control" (pp. 5–7). State regulations now specified the performance standards to which teacher evaluation must align, requirements for the inclusion of student growth data, training for evaluators, and many other outcome and procedural details. It must be noted, however, that these changes in Maine were not done in isolation but were influenced by nationwide pressures that accelerated in the late 20th and early 21st centuries.

National Pressures to Reform Evaluation and Supervision

Throughout the mid-1980s and 1990s, public scrutiny of education increased and statewide mandatory tests expanded (Donaldson, 2014). This focus on assessment and accountability continued into the 21st century. *No Child Left Behind* (the reauthorization of the Elementary and Secondary Education Act) passed both houses of Congress by wide margins (with bipartisan support) and was signed into law in 2002 by President George W. Bush (Urban & Wagoner, 2009). In its title and text the act called for proficiency for all students in all schools in all states, and for "a highly qualified teacher in every classroom" (Hazi & Ricinski, 2009, p. 2). The act also included significant penalties—reorganization or closure—for schools failing to meet the goal. Shortly after the *No Child Left Behind* act the National Governor's Association identified related goals: "define teacher quality, focus evaluation policy on improving teaching practices, incorporate student learning into teacher evaluation, create professional accountability through developing career ladders, train evaluators in pre-service

programs, and broaden participation in evaluation designs" (p. 2). Part of the strategy, then, for leading all students to proficiency was a focus on current and incoming teachers.

Calls for teacher supervision and evaluation changes were not new. For example, Weisberg et al (2009, p. 2) quote a 1936 issue of *The New York Times*: "Whether these incompetents were unfit to teach at any time, or have been rendered unfit by the passing years, is a matter of opinion. The question is, why are they allowed to remain?" In recent years, however, the rhetorical calls ultimately led to legislated and regulatory changes. By the late 2000s the United States public had years of data that were not encouragingly trending toward achieving the *No Child Left Behind* mandate of proficiency for all by 2014 ("NAEP 2008: Trends in Academic Progress," 2009). A global economic challenge impacted the end of George W. Bush's presidency and the beginning of President Barack Obama's first term. The Obama administration used a large pool of money, $4.35B made available through federal stimulus funds, to promote certain initiatives through the Race to the Top program (Marzano & Toth, 2013). Changes to teacher evaluation systems were among those requirements for states seeking funding (Darling-Hammond, 2013; Marzano & Toth, 2013).

Simultaneously, public attention toward teacher employment issues such as tenure and evaluation increased. In 2009 *The New Yorker* published "The Rubber Room" regarding New York teacher reassignment centers, followed by Garrett and Cegnar's documentary with the same title (2010). The 2010 film *Waiting for Superman* brought further attention to teacher evaluation, capturing significant public interest: "[T]his movie, along with a veritable flood of commentaries on local and national news shows, brought the issue of teacher evaluation into sharp relief…the inadequacies of teacher evaluation systems were well known and a matter of public discussion" (Marzano & Toth, 2013, p. 3).

Congress had not yet acted to reauthorize ESEA/NCLB, so the law of the land still called for proficiency for all students by 2014. The Obama administration applied pressure via another lever, using ESEA/NCLB "flexibility waivers" to allow states to redefine their path to proficiency providing that certain changes were made, including changes to teacher evaluation systems (Darling-Hammond, 2013). Through the Race to the Top initiatives and the ESEA/NCLB flexibility waivers, the administration moved many states to begin significantly revising teacher and principal evaluation systems to include such factors as student achievement scores, performance standards, and multiple categories of ratings (Darling-Hammond, 2013; Marzano & Toth, 2013). Darling-Hammond (2013) describes the subsequent state of affairs:

> Prodded by the requirements of Race to the Top grants and federal waivers from No Child Left Behind, states and districts across the United States have been changing their policies at a dizzying rate, often with little chance to consider the research base for practice. Most are trying to accommodate these new mandates while constructing productive, manageable systems that support high-quality practice, teacher learning, and student success. This is no easy feat. There are many landmines in this territory, and little available information that can offer decision-makers both research evidence and practical examples to inform this work. (vii)

The federal changes continued with the late-2015 passage of the *Every Student Succeeds Act* (ESSA), implementation of which is phased in through the 2017–2018 school year to replace fully *No Child Left Behind*, the *Race to the Top* initiatives, and the NCLB flexibility waivers (Posey, 2016). The only immediate PE & PG impacts of the federal ESSA in Maine were an optional extra year to pilot systems and a one-year delay of the requirement to use student growth data in PE & PG systems. Overall in Maine the 2012 through 2015 statutory and

regulatory changes to teacher evaluation and supervision, created partially in response to now-outdated federal pressures, remain in effect.

Maine's Current Requirements

Under the statutory authority of Title 20-A MRSA §13706, Maine Department of Education Chapter 180 (last amended March 18, 2015) details the regulatory requirements for school districts with regard to PE & PG systems. The regulation lays out details for system development (e.g., process, stakeholder involvement, governance groups), describes the process for Department approval of local plans, and defines relevant terms (e.g., instructional cohort, rating level). Chapter 180 identifies that PE & PG systems for teachers must align to the Interstate Teacher Assessment and Support Consortium (InTASC) Model Core Teaching Standards (2011), and pre-approves the alignment of several published models (e.g., Danielson, Marzano, Marshall).

Significant text in Chapter 180 is devoted to the required inclusion of student learning and growth measures in summative evaluation, describing and defining, for example, characteristics of permissible measures, limitations on attribution of student data to individual teachers, alignment to the Student Learning Objectives framework, and use of collective growth measures. Requirements for calculating the summative evaluation rating are described, along with consequences of summative ratings and description of subsequent Professional Improvement Plans or Professional Growth Plans. A peer review component is required and described, and many additional implementation details including training are outlined.

As a whole, the regulatory details appear to include limited and generally vague language regarding formative feedback and support (i.e., teacher supervision), with a strong focus on summative evaluation. For example, Chapter 180 (2015) includes 58 overall instances of

derivatives of the term "evaluate" (excluding repetitions of the title) and only 11 instances of derivatives of the terms "growth" or "supervise" for the explicit purpose of improving teacher practice (excluding repetitions of the title and excluding references to student growth measures). Local school districts may add supervisory components and supervisory resources beyond those required by law. Maine's statutory and regulatory structure, as described above, does not explicitly provide but potentially allows what Zepeda (2012) calls the opportunity "to link instructional supervision, professional development, and evaluation as a cohesive plan" (p. 13).

Maine school districts began the piloting and implementation stages for PE & PG systems over the course of several years, with some early adopters and others at or after the regulatory deadline. At the time data collection began approximately 150 school districts had submitted PE & PG system plans to the Maine Department of Education, and approximately 40% of those plans had received approval.[3] As many Maine school districts shifted from planning and piloting systems to implementation, I utilized the method described below and in Chapter 3 to contribute to the field and the literature.

Overview of Method

In this study I used a mixed-methods multi-site case study approach to describe the perceptions and experiences of teachers, evaluators, and supervisors in piloting and early local implementation of Maine's mandated multi-function systems of teacher performance evaluation and teacher professional growth. I purposefully selected eight sites (school districts) to yield a sample representative of Maine's school district implementation of PE & PG systems. The

[3] It is important to note that this approval rate did not necessarily indicate major regulatory problems with the remainder of the plans. In my personal experience with plan approval for my school district, several rounds of clarification and minor adjustment were needed to achieve final DOE approval.

sample included a rural and a non-rural site using each of the four teacher professional practice models widely used in Maine (Danielson, Marshall, Marzano, and the National Board for Professional Teaching Standards).

I collected data during the 2016–2017 school year; during this year most of the sampled sites were piloting their teacher PE & PG system district-wide. Thus, the data may not reflect what the systems would look like in a more advanced stage of implementation. At each site (school district) I conducted qualitative semi-structured interviews with people knowledgeable about the district's PE & PG system implementation: a teacher, an evaluator of teachers, and (if the district employed such a person) a non-evaluator supervisor of teachers. Thus, I interviewed two or three individuals at each site, for a total of 20 interviews in the study. Additionally, I collected quantitative data from 302 teachers, evaluators, and supervisors across the sites via an online survey.

I used these qualitative and quantitative data in combination to answer the research questions via within-case and cross-case analysis. I utilized Attribute Coding, Evaluation Coding, and Values Coding in the first cycle of qualitative data analysis, and Pattern Coding in the second cycle to condense codes into themes and concepts (Saldaña, 2016). In quantitative data analysis, I used descriptive and basic inferential statistics (paired sample t-test, Kruskal-Wallis test with post-hoc analysis) to describe practitioner perceptions and to determine if there were significant patterns in these perceptions across roles in the school, professional practice model, or rural status. I present the research findings within-case and cross-case, in a manner organized around the research questions and the conceptual framework.

List of Definitions

For the purposes of this study these terms are defined as follows:

Practitioners at the local level means teachers, supervisors of teachers (e.g., Instructional Coach), and evaluators of teachers (e.g., Principal). *Teachers* means those who describe their primary role in school districts as providing instruction to students as a professional teacher in Maine. The term includes, for example, such roles as classroom teacher, physical education teacher, and reading interventionist. *Evaluators* means those who describe themselves as one of the people in school districts responsible for directly evaluating teacher performance for employment purposes. The term includes, for example, such roles as principal and director. *Supervisors* means those who describe themselves as one of the people in school districts responsible for directly supporting teacher performance and growth, rather than evaluating that performance for employment purposes or providing direct instruction to students. The term includes, for example, such roles as literacy coach and instructional supervisor. Some school staff in Maine play multiple roles, which cause overlap between the functions of teaching, supervising, and evaluating. In this study, participants will self-report their primary function. *School Districts* or *Local* means the school district level (e.g., Regional School Unit, Municipal School District).

Teacher performance evaluation is a formal summative function focused on the organizational need for accountability, determining and documenting the level of a teacher's performance over a specific time period (Hazi & Rucinski, 2009; Platt et al., 2000).

Teacher professional growth is a formal or informal formative function involving both self-directed and guided teacher improvement; the process of guiding teacher professional growth is referred to as *supervision* (Glickman, Gordon, & Ross-Gordon, 2013).

Teacher effectiveness and *improved teacher effectiveness* in this study are as perceived and defined by the individual practitioners for the following reasons: (1) the term is not defined in Maine's PE & PG Systems regulation; (2) the local definitions, if they exist, would likely be unique to the professional practice model, locally-determined weighting of system components such as student growth data, and local values; and (3) even within the same school district educators' perspectives on effectiveness vary widely, drawing upon different logics and "intellectual, professional, and cultural histories" (Rigby, 2015, p. 378).

Significance of the Study

For local Maine practitioners, the research findings prove instructive about varied approaches to teacher evaluation and teacher supervision under the new Maine law, and about local factors that contribute to or arise as barriers to improved teacher effectiveness. For policymakers, the research findings prove instructive about local school district outcomes following this state action, and provide perspective about the actual impact of the law (in an early stage) relative to the intended impact. For researchers, the findings capture an important moment in time in Maine teacher evaluation and supervision, and provide the basis for further research as Maine shifts from piloting and early implementation to later implementation or refinement of these PE & PG systems.

Finally, while the results provide an in-depth look at the phenomenon and rich data for these eight sites in Maine, the nature of a case study design is such that the results are not generalizable. This study provides broad perspective for future studies examining state-wide PE & PG system implementation programs, particularly in states with geographic and demographic characteristics similar to Maine. I next expand on this overview by deeply exploring the theoretical and empirical literature regarding teacher supervision, teacher evaluation, multi-

function systems that attempt to address both functions, and the recent national and state initiatives.

CHAPTER 2

LITERATURE REVIEW

In Chapter 2 I review the empirical and theoretical literature relevant to this study about early local implementation of Maine's state-mandated Performance Evaluation and Professional Growth (PE & PG) systems for teachers. The literature reviewed in this section addresses major themes in teacher evaluation and supervision from the mid-to-late 20th century through current times, defining, describing, and clarifying distinctions between the two functions. The review shows tensions between evaluation and supervision, including misunderstanding of the meaning of supervision. Multi-function systems to concurrently address evaluation and supervision are described; these are directly relevant to this study as several such systems are pre-approved in Maine as professional practice models for teacher PE & PG systems. A well-regarded model leading to teacher professional growth is described as a basis for the theoretical and conceptual framework, and finally the emerging national and state research on teacher evaluation and supervision systems is summarized.

The teacher supervision and evaluation literature includes empirical and theoretical contributions. It is important to note in reading this review that values and choices are involved even in empirical pieces as researchers determine criteria for study. For example, when researchers study supervision or evaluation systems or elements of such systems in relation to student achievement, the chosen measure of student achievement (e.g., state reading or mathematics testing) is deemed important (e.g., Holtzapple, 2003; Jacob & Lefgren, 2008). In this regard, the empirical literature is weighted toward subjects frequently tested at the state level and weighted away from such topics as creativity and social studies.

Terms such as "teacher quality" and "teacher effectiveness" are also value-laden, and are often defined loosely and in varying ways (if at all). Maine's PE & PG regulation (Chapter 180, 2015), for example, does not define effectiveness directly, though one could infer that effectiveness in that context means acceptable performance relative to the pre-approved teaching standards and gains in student growth data goals. In the literature the terms tend to include an impact on student learning, often as measured using standardized assessments, but also tend to include a variety of characteristics such as the ability to teach all learners (as opposed to raising mean scores), the ability to inspire, and the ability to address social or emotional needs of students. As detailed in Chapter 3, in this study the definition of "effectiveness" was left to each interview or survey participant.

It is further important to note that many authors blend empirical and theoretical elements in their work. For example, Danielson describes the research base for her *Framework for Teaching* (2013) as beginning with research synthesized by the Educational Testing Service and then evolving since 1996 through the recommendations of professional organizations, through additional research, and through efforts to connect the framework to outside efforts (e.g., the Measures of Effective Teaching project and the Common Core State Standards). Darling-Hammond (2013) provides a similar description, drawing on empirical research from a number of peers but also informed and influenced by feedback from a variety of professional organizations and from panels of experts. Marzano (2013) describes the most empirical approach of the major systems designers, citing decades of research informing the development of his teacher evaluation model and ongoing action research to evaluate the model. In contrast, Marshall (2013) takes a more theoretical approach, informed by the empirical literature but candidly drawing from his experiences as a principal and as a consultant. Weisberg et al. (2009)

also bring an approach mixing theory with study (not peer reviewed), coming from an advocacy standpoint. As a whole, most contributions to the literature in this field are primarily theoretical; throughout this review I aid the reader by specifically noting empirical contributions.

Regarding specific aspects of this literature review, five main types of literature are reviewed: 1) teacher evaluation, 2) teacher supervision, 3) multi-function systems of evaluation and supervision, 4) Glickman, Gordon, and Ross-Gordon's *SuperVision* (2013), a major leg of the theoretical framework, and 5) the emerging research regarding new systems in this era of federal and state mandates. It is important to make these distinctions in order to understand how educators currently think about combined models of supervision and evaluation, although in theory they are distinct. Thus, understanding how supervision and evaluation are combined in practice was one of the purposes of this study. In the next section I discuss evaluation and supervision separately before describing theoretical and for-profit approaches that attempt to combine the two functions.

Teacher Evaluation and Teacher Supervision

In this field's literature, and to a greater extent with practitioners and laypeople, the terms *evaluation* and *supervision* are frequently and incorrectly used as synonyms. Hazi and Ricinski (2009) note tension between evaluation and supervision in the literature dating at least back to 1920. As roles and interactions in schools became more formalized and public scrutiny of education increased, formal evaluation seems to have dominated supervision to the point where the two are "forever entangled" (p. 14) and now "supervision is usually understood as teacher evaluation in the schools" (p. 2). In a study of 100 teachers and their principals Ponticell and Zepeda (2004) found that "for all teachers and for the vast majority of principals, supervision was, quite simply, evaluation" (p. 47). The purposes of evaluation and supervision, however, are

distinct; evaluation is primarily an assessment of performance while supervision is ongoing support for professional growth purposes. In the following two sub-sections I delve deeply into each topic, defining and describing *evaluation* and *supervision* and clarifying distinctions between the two functions.

Teacher Evaluation

Teacher evaluation (i.e., summative evaluation) focuses on the organizational need for accountability, determining and documenting the level of a teacher's performance over a specific time period (Hazi & Rucinski, 2009; Platt et al., 2000). Olivia and Pawlas (2004) defined summative evaluation as "assessment of teacher performance by an administrator for the purpose of making decisions about such matters as tenure, retention, career ladder, and merit pay" (p. 367). Teacher evaluation fulfills this personnel function through a process Platt et al. (2000) further describe:

> Evaluation – means the process of defining goals and identifying, gathering, and using evidence…to improve professional performance and to assess job effectiveness by determining whether the facts substantiated by the evidence that has been documented satisfy the school district's performance standards. (p. 143).

Evaluation is typically formal, using standardized criteria, forms, and procedures defined in policy. Teacher evaluation cycles are time-bound, with summative evaluations occurring typically annually for probationary teachers, and typically once every one to three years for teachers who are tenured or on a continuing contract. An evaluator (often a principal, assistant principal, or director) designated by the superintendent (Platt et al., 2000) gathers evidence from one or more sources (e.g., classroom observations) during that cycle. The evaluation cycle may also include components such as goal-setting and self-assessment. At the conclusion of the cycle

an evaluation report, often referred to as the summative evaluation, is documented. If the summative evaluation determination is satisfactory, a new cycle typically begins, while an unsatisfactory summative evaluation may lead to development of an improvement plan or employment action such as dismissal.

While other factors (e.g., socioeconomic status, attendance) influence student outcomes, it is generally accepted that teacher quality is an important factor in student outcomes (Goldrick, 2002). Thus, evaluating teacher quality (in addition to improving teacher quality via supervision) is viewed as an important function in school decision-making such as offering contracts, renewing contracts, and determining positive or negative consequences for teachers (Koppich, 2005; Medley & Coker, 1987). This evaluation function, however, is not easily done well as evidenced by a literature dominated for decades by critiques and recommendations for change.

Teacher evaluation systems and their implementations receive considerable attention from many directions. Evaluation is "flawed, contested, and problematic" (Hazi & Rucinski, 2009, p. 3). In the literature significant criticism of evaluation systems dates back at least to a 1984 RAND study (Darling-Hammond, 2013; Marzano & Toth, 2013), but public attention has increased in the past decade. Marzano and Toth (2013) note three recent works which have intensified public scrutiny: *Rush to Judgment* (2008), *The Widget Effect* (2009), and especially the film *Waiting for Superman* (2010). The researchers note, "This movie, along with a veritable flood of commentaries on local and national news shows, brought the issue of teacher evaluation into sharp relief" (Marzano & Toth, 2013, p. 3). Darling-Hammond (2013) notes that, relative to the early 1980s, "the broad landscape for teacher evaluation has changed little, and impatience with the results of weak systems has grown" (p. 2).

In the paragraphs below I describe the ways numerous authors critique and suggest improvements for teacher evaluation. These assertions in the literature emerge from research and theoretical origins. It is important to note that the extent of and responsibility for evaluation system claims may vary across authors (e.g., some authors are more critical of system design while others are more critical of union protections).

Teacher evaluation could be improved through stronger system design and greater evaluator consistency. First**,** some authors critique that evaluation is not based on a shared understanding of good teaching practices (Darling-Hammond, 2013; Marshall, 2013; Platt et al., 2000) and staff lack a common technical language for discussing teaching practice (Spillane, Reiser, & Reimer, 2002). This lack of shared language and understanding impacts effective dialogue among teachers, between teachers and administrators, and among administrators. "Administrators have wide-ranging definitions of what constitutes acceptable and excellent teaching" (Platt et al., 2000, p. 32). Further, evaluators—typically administrators—are insufficiently trained (Darling-Hammond, 2013; Marshall, 2013; Platt et al., 2000; Weisberg, Sexton, Mulhern, & Keeling, 2009), may lack content expertise, and often disagree with each other. In a study of 46 principals and 322 teachers, Medley and Coker (1987) noted a low (0.20) correlation between principal ratings of teachers and direct measurements of teacher effectiveness. Platt et al. (2000) note that "[t]he grade or ranking a teacher gets depends on who the evaluator is and what the evaluator believes…unions rightfully point to the discrepancies and complain that the system is unfair" (p. 32). In recent decades, some gains have been made in connecting evaluation ratings to student gains (Holtzapple, 2003), but the relationship is far from consistent (Marzano & Toth, 2013).

In teacher evaluation there is tension between system rigidity and protection from arbitrary action, influenced by a desire for fairness (DeSander, 2000) and legal defensibility (Glickman, Gordon, & Ross-Gordon, 2013). Koski (2012) further describes this tension:

> From the teacher's or union's perspective, such rigid and predictable systems minimize arbitrary administrative evaluations that may be based on personality conflicts, inappropriate factors, or sloppy observations. From the administration's perspective, these rigid systems do not sufficiently account for a teacher's impact on student performance, are time consuming due to paperwork, and do not give the principal sufficient flexibility to provide meaningful feedback and, ultimately, make decisions about the teacher's future employment. (p. 86)

Similar to Koski, Darling-Hammond (2013) and Glickman et al. (2013) note that rigid systems may preclude tailoring helpful feedback to teacher needs. Evaluators limit potentially helpful or challenging feedback due to fear of litigation (DeSander, 2000; Glickman et al., 2013) and a "norm of non-interference" with peers (Goldstein, 2007, p. 481). Sullivan and Zirkel (1998) found that a minority of states protect teacher evaluation records from public disclosure; many evaluators must consider that the public may see their evaluative comments following an open records or freedom of information request. "Administrators walk a veritable minefield of legal issues when attempting to satisfy the demands of staff, students, unions, and the public" (Sullivan & Zirkel, 1998, p. 367).

Teacher evaluation could be improved through better data collection. Authors frequently critique the quantity and quality of observational data used in teacher evaluation. The number and length of classroom observations in a typical evaluation cycle makes up a small fraction of a teacher's work with students (Darling-Hammond, 2013; Marshall, 2013; Platt et al.,

2000; Weisberg et al., 2009), and observations are often limited by the use of narrow checklists or by an attention-consuming requirement to script the entire observed lesson (Darling-Hammond, 2013; Marshall, 2013, 2014; Platt et al., 2000). This paucity of data is one factor leading evaluators to doubt themselves, possibly contributing to a tendency for principals to give more generous ratings than full-time consulting teachers (Goldstein, 2007). Weisberg et al. (2009) found in their study that 64 percent of tenured teachers were observed no more than twice during their most recent evaluation cycle, totaling 75 minutes on average. The rates were little different for probationary teachers—e.g., 59% observed no more than twice for an average total of 81 minutes. While Maine was not part of the Weisberg et al. study, as a point of comparison a Maine teacher would likely provide over 45,000 minutes of instruction in a school year. To observe a Maine teacher for even 1% of their instruction with students would require at least 450 minutes (7.5 hours) of observation annually.

In terms of validity, a classroom observation - one of the major sources of information in a typical evaluation system - provides perspective about a single lesson, without the context of the unit of study or assessment of that learning (Marshall, 2013). The pre-conference model often used allows teachers to prepare for an observation in a way different from their normal day-to-day preparation (Hazi & Rucinski, 2009; Holtzapple, 2003; Marshall, 2013), and the presence of the evaluator in the classroom can change how students and the teacher act (Marshall, 2013). Sampling error occurs in observations when the lesson observed, for whatever reason, does not represent a teacher's typical behavior (Marzano & Toth, 2013). Finally, the use of some observation instruments (e.g., checklists expecting certain behaviors in every lesson) may lead to perverse consequences such as lessons designed to meet the checklist rather than to meet student needs (Darling-Hammond, 2013).

Further, data collected in many evaluation systems do not directly address the desired outcome; evaluations tend not to include evidence of student learning (Darling-Hammond, 2013; Marshall, 2013; Marzano & Toth, 2013), which some believe should be significant criteria in evaluation. Many authors in the teacher evaluation literature include teacher analysis of student data in evaluation. A smaller set of authors promotes the direct use of student scores in determining the summative evaluation rating: "Presumably, the best teacher evaluation systems would provide evidence of teaching through the documentation of learning—both student learning and teacher learning" (Wilkerson, 2000, p. 180). Problems are noted with direct inclusion of student scores; when authors discuss inclusion of student scores in teacher evaluation they generally promote the use of multiple, carefully constructed measures (Darling-Hammond, 2013; Marzano & Toth, 2013).

The quantity and quality of observational data reflects that an evaluator's time is often inadequate to effectively evaluate a large number of teachers (Darling-Hammond, 2013; Goldstein, 2007; Marshall, 2013; Platt et al., 2000). Darling-Hammond (2013) notes that many evaluations tend to be perfunctory, possibly influenced by the scarcity of time as principals and other evaluators manage a greater evaluative load than typical in business—along with many other responsibilities. In a survey, Weisberg et al. (2009) found that nine percent of teachers had missed their most recent evaluation. Moreover, Platt et al. (2000) point out, "Generally, administrators in American schools supervise and evaluate too many people annually to be able to do a creditable job" (p. 24).

Teacher evaluation could be improved through more detailed and accurate ratings. Whether influenced by insufficient observational data or some other factor, evaluation ratings tend not to distinguish greatly between levels of teaching (Darling-Hammond, 2013; Marshall,

2013; Weisberg et al., 2009). Contributing to this challenge are the use of binary ratings (satisfactory or unsatisfactory) which leave little room for documented improvement (Marshall, 2013), and rating scales that do not provide sufficient description of teaching at various levels (Marshall, 2013; Marzano & Toth, 2013). Furthermore, evaluation ratings may hide instructional shortcomings by averaging them with responsibilities such as committees and coaching - a "leniency effect" or "halo error" (Platt et al., 2000, p. 11). According to Platt et al. (2000), "[A]lmost inevitably, teachers whose classroom performance is mediocre but who sustain the extracurricular or social life of the school will have received good to excellent evaluations" (p. 24). Weisberg et al. (2009) found in a study of 12 school districts that roughly 60% of teachers and administrators believe their "district is not doing enough to identify, compensate, promote and retain the most effective teachers" (p. 6).

Evaluation ratings tend to greatly exceed what one would expect; teachers generally expect and receive high ratings (Marshall, 2013; Marzano & Toth, 2013; Weisberg et al., 2009). Weisberg et al. (2009) found that more than 96% of teachers rated their own performance as a seven or higher on a 10-point scale. Bradshaw (2002) found in a survey of North Carolina teachers that 90% of respondents reported receiving performance ratings of "above standard" or "consistently superior" in a region with "continuing reports of less than stellar student performance" (p. 120). Weisberg et al. (2009, p. 6) found similar results; teachers received "uniformly positive evaluation ratings" even in schools where students struggled with academic standards (less than one percent of teachers received unsatisfactory ratings, and 94% received one of the top two ratings). Despite the high ratings a majority of teachers (58%) and administrators (81%) say there is a tenured teacher performing poorly in their school; 43% of

teachers say there is a tenured teacher in their school "who should be dismissed for poor performance"; Weisberg et al. found a "scarcity of formal dismissals" (p. 6).

Viewed through another lens, when Weisberg et al. (2009) compared teachers' perceptions of poorly-performing colleagues to the percentage who actually received unsatisfactory ratings, the differences were striking (e.g., 8% perceived vs. 1% actual in Chicago, 5% vs. 0% in Akron). Platt et al. (2000) describes supervisor fears which may lead to higher-than-expected ratings: harming relationships, eroding collegiality, losing professional friendships, seeming too demanding, losing good will, creating disruption, provoking an anger response. Marshall (2013) echoes that principals fear eroding relationships and losing cooperation, adding that teachers can subtly retaliate in many ways and that principals may lack practice with challenging conversations. Weisberg et al. (2009) summarize the challenge as a "vicious cycle":

> Administrators generally do not accurately evaluate poor performance, leading to an expectation of high performance ratings, which, in turn, causes administrators to face stiff cultural resistance when they do issue even marginally negative ratings. The result is a dysfunctional school community in which performance problems cannot be openly identified or addressed. (p. 23)

Further complicating the issue, teacher evaluation ratings are influenced by a number of factors outside of the teacher's control (Darling-Hammond, 2013; Hanushek & Rivkin, 2006; Holtzapple, 2003; Marzano & Toth, 2013). Examples that may buoy or decrease a teacher's demonstrated effectiveness include the effectiveness of the school or school and district leaders, the curriculum, the instructional materials, supportive specialists, and supportive peers.

"Observed schooling situations represent the outcomes of several interrelated choices – those of parents, teachers, administrators, and policymakers" (Hanushek & Rivkin, 2006, p. 23).

Teacher evaluation could be improved through greater impact on employment outcomes. As a whole, evaluation is widely perceived to have little impact on teaching and learning, except in egregious cases (Bradshaw, 2002; Darling-Hammond, 2013; Glickman et al., 2013; Marshall, 2013; Weisberg et al., 2009). Authors note that various parties are typically perceived responsible for this outcome: politicians and legislatures for excessive tenure and due process protections (Koski, 2012; Platt et al., 2000; Weisberg et al., 2009), teacher unions for protectionism and restrictive contracts (Koski, 2012; Platt et al., 2000), administrators for poor documentation of inadequate performance (Weisberg et al., 2009) or avoiding confrontation (Platt et al., 2000), teachers who don't challenge themselves professionally (Platt et al., 2000), and school systems for lack of support for struggling teachers (Darling-Hammond, 2013; Weisberg et al., 2009). Weisberg et al. (2009) found in a study of 12 school districts that teacher performance information is "almost exclusively used for decisions related to teacher remediation and dismissal," excluding other areas such as professional development, compensation, granting tenure, retention, and layoff order (p. 4). With these limits there are few, if any, extrinsic benefits to improving practice and receiving a higher evaluation rating.

Regarding dismissal of ineffective teachers, little impact is noted. Weisberg et al. (2009) found that "at least half of the districts studied have not dismissed a single non-probationary teacher for poor performance in the past five years" (p. 6). The highest annual dismissal rate found in the study was 0.04% in Cleveland. A large majority (86%) of administrators in the study "say they do not always pursue dismissal even if it is warranted" (p. 17); choosing not to

invest considerable time in a cumbersome process for an uncertain outcome. "Procedural errors still result in the reinstatement of teachers" (Sullivan & Zirkel, 1998, p. 379).

In the same study, districts were unlikely to "non-renew" even probationary teachers; over a five-year period the highest percentage of pre-tenure non-renewals found in a district was 3.1% (Weisberg et al., 2009). Platt et al. (2000) gives three reasons for this tendency to offer tenure to a high percentage of probationary teachers: supervisors become invested in their staff, prefer to stick with the known rather than take their chances with a new hire, and prefer to avoid the time investment in posting and hiring a position.

Unintended consequences of evaluation. All together, the evaluation process can result in negative side-effects in schools. Evaluation is often perceived to be unhelpful in improvement and primarily for punitive purposes (Darling-Hammond, 2013; Weisberg et al., 2009). The process can inadvertently discourage improvement through negative feelings, reduced participation, decreased willingness to alter behavior, and erosion of trust (Glickman et al., 2013; Marshall, 2013). Teachers who perceive themselves as performing well are maddened when colleagues they perceive as poor performers receive evaluation ratings similar to their own (Weisberg et al., 2009). Conversely, when dismissal of an ineffective teacher is sought the action may create widespread morale challenges, as staff "circle the wagons" to protect a peer (Marshall, 2013; Platt et al., 2000), "even if they know the teacher in question is incompetent" (Marshall, 2013, p. 31).

Calls for substantial reform in evaluation. In the preceding paragraphs I described criticisms and recommendations commonly noted with teacher evaluation. It is important to note that the overall tone of criticism in the literature does not reflect a lack of value for evaluation; in contrast, the improvement-seeking focus reflects the important potential authors see in

evaluation. While the exact nature or source of evaluation system challenges may be debated, many agree that significant room for improvement exists:

> The teacher evaluation process that has taken place, in many instances, has served no constructive purposes for improving performance. McLaughlin (1990, p. 403) points out, "Teachers are evaluated by one means or another in virtually every school district. And in most of those districts, teachers and administrators agree that the activity is ritualistic and largely a waste of time." (Wilkerson, Manatt, Rogers, & Maughan, 2000, p. 180).

Marshall (2013, p. 124) continues:

> In a more benign era, this ritual didn't attract much attention. Some teachers may even have appreciated the fact that it kept principals busy and out of mischief. But now the stakes are higher. Educators are being held accountable for the achievement of every single child. We can't avoid the conclusion that if our approach to end-of-year teacher evaluation is chewing up large amounts of time and rarely producing improvement, it needs to be changed.

Finally, authors note that evaluation is not sufficiently paired with professional development and support (Darling-Hammond, 2013; Weisberg et al., 2009). Weisberg et al. (2009) write that the inability to accurately assess instructional practice and act on that information prevents schools from providing differentiated professional development to ineffective teachers and "supporting growth among the broad plurality of hard-working teachers who operate in the middle of the performance spectrum" (p. 2). As noted above, supervision—"the helping or teacher professional development function" (Hazi & Ricinski, 2009, p. 4) — became dominated by "the personnel function" (p. 143) as roles and interactions became more formalized and as public scrutiny increased. In the next sub-section I describe teacher

supervision, followed by a section synthesizing authors' recommendations for multi-function systems that perform both supervision and evaluation functions.

Teacher Supervision

Teacher supervision (distinct from summative evaluation) is focused on ongoing feedback, teacher improvement, and teacher professional growth; the term *supervision* is broadly and inconsistently defined. Supervision may include a wide variety of formal and informal efforts, e.g., the clinical supervision model of conferencing before and after an observed lesson, staff development activities (large group, small group, or individual), portfolio development, goal-setting, data analysis, and reflection (Ponticell & Zepeda, 2004; Eady & Zepeda, 2007; Hazi & Ricinski, 2009). Similarly broad, authors note that supervision may be provided by a wide variety of people. Typical titles of those providing supervision include: assistant superintendent, consultant, curriculum coordinator, curriculum director, department head, instructional coach, instructional specialist, mentor, principal (and/or assistant principal), specialist, staff developer, or team leader (Beach & Reinhartz, 1989; Oliva & Pawlas, 2004; Wiles & Bondi, 2004).

Definitions of supervision. Teacher supervision is formally defined in a variety of ways by different authors; confusion regarding role definition has been present since at least 1975 (Wiles & Bondi, 2004). Arriving at a single definition is difficult with varied priorities about teaching and learning and as the supervisory role varies greatly from location to location (Oliva & Pawlas, 2004). Beyond Hazi and Ricinski's description of teacher supervision as "the helping or professional development function" (2009, p. 4), other major authors describe supervision as:

- "assistance for the enhancement of teaching and learning" (Glickman et al., 2013, p. 9)
- "enhancing teacher thinking, reflection, and understanding of teaching" (Ponticell & Zepeda, 2004, p. 43)

- “a means of offering to teachers, in a collegial, collaborative, and professional setting, specialized help in improving instruction and thereby student achievement" (Oliva & Pawlas, 2004, p. 11)
- “working with teachers to increase the quality of student learning through improved instruction” (Beach & Reinhartz, 1989, p. 8)

Additional definitions exist in the literature, with common themes of helping or assisting teachers in order to enhance the service they provide to students. Fundamentally, supervision should be a teacher-focused process guided by principles of adult learning (Ponticell & Zepeda, 2004).

While the ultimate purpose of supervision—indirectly helping students—seems evident, peeling back another layer reveals more complexity:

> The goal of supervision is to improve instruction. It sounds nice, until we ask for a definition of what type of instruction we wish to improve. Effective teaching, to a large extent, depends on what you are trying to teach. Different instructional goals require different teaching strategies. (Glickman et al., 2013, p. 81).

This emphasizes the situational nature of supervision: the appropriate service is dependent on factors including the instructional goals, the strengths and needs of the teacher(s), the career stage of the teacher, and organizational goals (Beach & Reinhartz, 1989; Wiles & Bondi, 2004; Glickman et al., 2013). As shown in a brief historical overview, the role of supervision has also changed relative to trends and needs in schooling and in the nation.

Historical and current approaches to supervision. Supervision has been linked to school and school district administration largely since its inception; as districts composed largely of one-room schoolhouses increased in size, roles such as principal teacher and assistant

superintendents were added (Oliva & Pawlas, 2004; Wiles & Bondi, 2004). Early in American education, those in such roles typically approached supervision from a perspective of telling, directing, and judging (Wiles & Bondi, 2004). Wiles and Bondi (2004) note how shifts in American social and economic conditions over the decades impacted the implementation of supervision in American schools. When efforts regarding efficiency and division of labor impacted industry, supervision focused on specialization and regulations. Early 20th century expansions in knowledge about psychology led to a testing focus, and World War II and Sputnik were associated with a focus on technical proficiency and prescriptive curriculum.

Then, in the mid-1960s to early 1970s research on teaching and learning coincided with supervision shifts to a more clinical model focused on professional dialogue about classroom events (e.g., pre-conference, observation, post-conference, analysis), and funding cuts in the late 1970s and early 1980s concurred with a return to basics (Beach & Reinhartz, 1989; Oliva & Pawlas, 2004; Wiles & Bondi, 2004). Simultaneously, as testing, evaluation, and unionism grew, supervision began to overlap more with management roles, and through the 1980s and 1990s new supervisory titles emerged along with new technologies and new laws (Wiles & Bondi, 2004). Oliva and Pawlas depict these major periods in a similar fashion (Table 2.1, below).

Over time, a strong shift is noted from a directive view of supervision to a focus on human relations, group dynamics, and the democratic process (Oliva & Pawlas, 2004). Along the same lines, some authors (Beach & Reinhartz, 1989; Glickman et al., 2013) categorize approaches as directive, collaborative, or non-directive. With directive approaches, supervisors generally approach their work from a prescriptive perspective focused on inspection

Table 2.1. *Major Periods in the Historical Development of Supervision*

Period	Type of Supervision	Purpose	Person Responsible
1620 – 1850	Inspection	Monitoring rules, looking for deficiencies	Parents, clergy, selectmen, citizens' committees
1850 – 1910	Inspection, instructional improvement	Monitoring rules, helping teachers improve	Superintendents, principals
1910 – 1930	Scientific, bureaucratic	Improving instruction	Supervising principals, principals, general and special central-office supervisors, superintendents
1930 – 1950	Human relations, democratic	Improving instruction	Principals, central-office supervisors
1950 – 1975	Bureaucratic, scientific, clinical, human relations, human resources, democratic	Improving instruction	Principals, central-office supervisors, school-based supervisors
1975 - 1985	Scientific, clinical, human relations, human resources, collaborative/collegial, peer/coach/mentor, artistic, interpretive	Improving instruction, increasing teacher satisfaction, expanding students' understanding of classroom events	Principals, central-office supervisors, school-based supervisors, peer/coach/mentor
1985 – present	Scientific, clinical, human relations, human resources, collaborative/collegial, peer/coach/mentor, artistic, interpretive, culturally responsive, ecological	Improving instruction, increasing teacher satisfaction, creating learning communities, expanding students' classroom events, analyzing cultural and linguistic patterns in the classroom	School-based supervisors, peer/coach/mentor, principals, central-office supervisors

Note. Reprinted from Oliva & Pawlas, 2004, p. 5

and control. Ponticell and Zepeda (2004) note conflict between this approach and the idea that an adult also needs to be self-directing. A non-directive supervisory approach, maximizing teacher responsibility, may for example focus on teacher development of a plan with minimal supervisory responsibility (Glickman et al., 2013). The overall strongly recommended style in the contemporary literature is a collaborative and congenial approach focused on reflective inquiry, in which teachers are central actors, in which supervision is tailored to the individual, and in which teachers and supervisors in a trusting relationship together assume responsibility for instructional improvement (Oliva & Pawlas, 2004; Zepeda, 2006; Zepeda, 2012; Glickman et al., 2013).

Contemporary models of supervision. Along with connecting to different teaching strategies, authors detail different supervisory strategies and roles to address various situations (Oliva & Pawlas, 2004) and varying teacher needs (Beach & Reinhartz, 1989; Zepeda, 2002). Several such models are described here. Oliva and Pawlas (2004), for example, break supervision down into the following types: administrative, clinical, collaborative, consultative, developmental, differentiated, educational, general, instructional, and peer supervision. Without minimizing more detailed facets, Wiles and Bondi (2004, p. 12) categorize supervision into six functions (supervision as an act of administration, as an act of curriculum work, as an instructional function, as an act of human relations, as management, and as a generic leadership role). Oliva and Pawles (2004) similarly frame three supervisory domains (Figure 2.1, below), with each connecting to four roles for a total of twelve supervisory role-domain connections (e.g., coordinating instructional development, coordinating curriculum development, evaluating staff development). As shown in the same figure, Oliva and Pawlas (2004) build these supervisory domains and roles on a foundation of theory and skills.

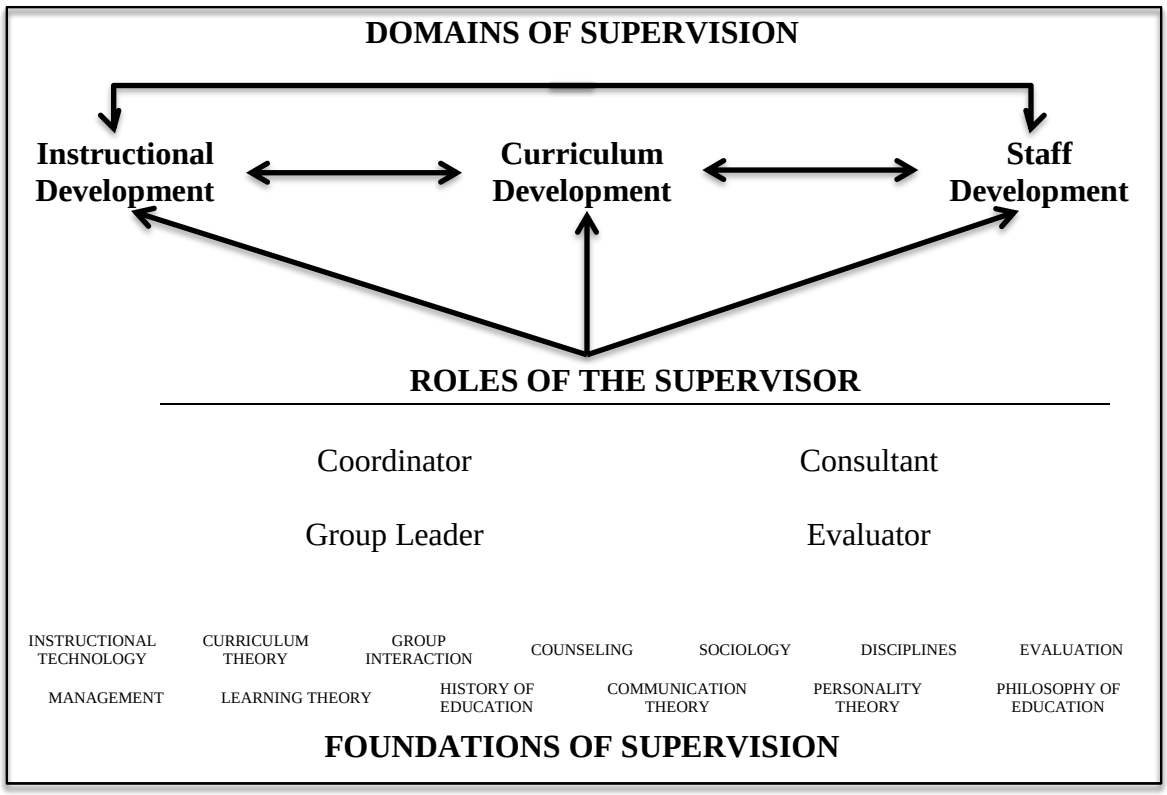

Figure 2.1. Oliva and Pawlas' conceptual model of supervision (2004, p. 21).

Glickman, Gordon, and Ross-Gordon (2013) identify three categories of similar "prerequisites": knowledge, interpersonal skills, and technical skills (p. 14). In this depiction, supervision ("SuperVision," as shown in the theoretical framework at the conclusion of this chapter) flows through technical and cultural tasks before uniting individual goals with school and community goals. Ultimately, all streams point to the outcome of improved student learning.

Supervisor expertise—knowledge and technical skills—are regarded as important: "Numerous studies have shown that if teachers seek help at all—and many do not—they want the supervisor to demonstrate the type of expertise that will help them in the classroom" (Oliva & Pawlas, 2004, p. 46). However, without minimizing the value of knowledge and technical skills, authors emphasize the importance of interpersonal skills. Oliva and Pawlas (2004) continue, "They also expect the supervisor to be able to develop rapport with them and work

with them in a cooperative way" (p. 46). In analyzing styles of interaction between supervisors and teachers, a collegial relationship is typically preferred (Glickman et al., 2013), though in some situations there is benefit to a superior-subordinate relationship, such as when there is need to maintain psychological distance from teachers (Oliva & Pawlas, 2004). A supervisor's total responsibilities may influence when this separation is needed. Oliva and Pawlas (2004) note that a supervisor purely in a helping role, such as a clinical supervisor, has little need for psychological distance, while a supervisor who also has responsibilities such as discriminating between teachers and making decisions about retention or dismissal may need greater psychological distance. This connects to a large and perpetual discussion in the literature regarding the links between supervision and administration; "the question of whether [supervisors] should be part of management is…a storm center among specialists in supervision" (p. 4).

Role tensions and other challenges in supervision. Tension is noted between the desired collaborative, trusting relationship and conflicting functions when the supervisor is also an administrator (with responsibilities such as summative evaluation, resource allocation, and employment decisions). Oliva and Pawlas (2004) summarize this ongoing discussion in the literature, while recognizing that in small schools and small school districts resources and staff sizes are such that administrators must also be supervisors. Wiles and Bondi (2004) also note that collective bargaining has pushed the field of supervision into a management posture. Noting that "every administrator is ipso facto a supervisor" to at least some degree (p. 14), Oliva and Pawlas share a continuum of basically full-time administrators, administrators who supervise part-time, supervisors who administrate part-time, and full-time supervisors. Staff supervisors, without administrative line authority, do not have as much power. This can pose challenges to

implementing change and enforcing policy (Beach & Reinhartz, 1989) but also may reduce interference with a positive teacher response by avoiding a threatening environment (Oliva & Pawlas, 2004). Line supervisors, with more authority, can fall back on power and status if needed, but likely will have difficulty creating and maintaining rapport and trust with teachers, particularly if they are responsible for providing evaluative ratings. Further, busy administrators may have difficulty devoting sufficient time to helping teachers.

Time limitations and tensions for those in dual administrative and supervisory roles are not the only challenges facing the field of teacher supervision. Other resources, such as funds for staff development, may be scarce (Eady & Zepeda, 2007). Time pressures and reduced flexibility pose challenges as states increase the number of mandates for testing, curriculum, and teaching; these mandates also potentially create "intrarole conflicts" for supervisors tasked with reforms with which they disagree (Oliva & Pawlas, 2004, p. 13). Continued diverse conceptions of supervision and of good teaching make the supervisory process difficult (Oliva & Pawlas, 2004). Finally, supervisors face a challenge of moving from helper to judge, as they are frequently tasked with playing a role in teacher evaluation (Beach & Reinhartz, 1989). Despite the challenges noted, supervision is widely regarded as important in the literature (Beach & Reinhartz, 1989; Oliva & Pawlas, 2004; Wiles & Bondi, 2004; Glickman et al., 2013) and may be increasingly important as schools manage a workforce with a lack of qualified applicants and high attrition (Zepeda, 2006). Such challenges may be particularly fierce in rural communities.

Special Challenges for Rural Teacher Evaluation and Supervision

Authors in the literature on rural education have described consistent difficulties facing rural schools; some of these challenges directly impact teacher evaluation and supervision. Recruiting and retaining professional staff is a long-standing challenge for rural and small

schools (e.g., Helge & Marrs, 1981). "Administrators in rural or small schools find it extremely difficult to locate and hire qualified teachers who are likely to fit in smoothly to both the school and community and to stay in that job for a long time" (Lemke, 1994, p. 20). More emphasis in the rural education literature is placed on attracting and retaining staff than on stringent evaluation: "[W]hen experienced teachers are dismissed on the basis of unsatisfactory evaluations and inadequate supervision and staff development, they are difficult to replace" (Eady & Zepeda, 2007, p. 6).

Factors contributing to the staffing challenge include but are not limited to geographic and social isolation, expanded responsibilities and roles relative to larger schools, the importance of fitting in with local cultural norms, and generally lower pay than at larger or urban districts (Helge & Marrs, 1981; Lemke, 1994; Schwartzbeck & Prince, 2003; Jimerson, 2005). The lower salaries at rural schools are tied to funding resources; many rural districts are in financial distress associated with older populations, declining population, and lack of property tax base (Schwartzbeck & Prince, 2003; Jimerson, 2005). Further, due to small student and staff populations teachers at rural schools often have a more broad set of responsibilities relative to teachers at larger schools, such as more non-instructional duties (which limits planning time) and responsible for multiple grades or multiple subjects (Schwartzbeck & Prince, 2003; Jimerson, 2005).

Addressing these situational needs, Lemke (1994) asserted that "the 'ideal' rural teacher is certified to teach more than one subject or grade level, can teach students with a wide range of abilities in the same classroom, is prepared to supervise extracurricular activities, and can adjust to the community" (p. 19). Lemke emphasizes the community differences when speaking about teacher induction. When a teacher newly begins at any school, induction to the role, the

curriculum, and the school culture are necessary. At a rural school in a small community the relative anonymity of living in a larger area does not exist and teachers face increased scrutiny, therefore promoting community fit is an important part of teacher recruitment and induction (Lemke, 1994).

Supervision and staff development at small rural schools, though, often rely on fewer people than at larger school districts. Eady and Zepeda (2007) point out that "[m]uch of the time, a single administrator must adequately supervise staff, evaluate staff, and provide staff development" (p. 6). This single point of contact brings into full focus the evaluator-supervisor role tensions described in the literature on supervision, perhaps exacerbated by the closeness and frequency of work together demanded of a very small staff.

In providing professional development the geographic isolation of rural communities exacerbates the funding scarcity, adding additional cost such as mileage and lost time for consultants or participants. Jimerson (2005) notes that "if professional development is only available in centralized settings, the distance can be a disincentive for rural teachers to participate" (p. 213). In their study of rural middle school principals Eady & Zepeda (2007) found similar difficulties in providing supportive opportunities for staff development, due to distance from academic communities. Further, constant staff turnover makes long-range planning for staff improvement "virtually impossible" (Helge & Marrs, 1981, p. 4), directly or indirectly impacting student engagement and achievement by disrupting staff cohesion (Ronfeldt et al., 2013).

Finally, relevant to Maine's requirement to include Student Learning Objectives in Performance Evaluation and Professional Growth systems, small group sizes add to test score volatility (Jimerson, 2005). The issues that Coladarci (2003) found with small school score

volatility relative to NCLB's *Adequate Yearly Progress* may emerge at the individual teacher level as districts implement Student Learning Objectives. "The problem is that when a small school drops below the AYP target one year, it is quite likely that this school had a 'bad bounce' rather than a real decline due to factors such as weak instruction, poorly aligned curriculum, or ineffective leadership'' (2003, p. 2). In Chapter 3 I will describe how I designed the study to include the perspectives of both rural and non-rural Maine schools.

In the preceding sections I reviewed the literature on teacher evaluation and teacher supervision separately, highlighting the differences between these functions. Some authors have created or proposed multi-function systems intended to address supervision and evaluation needs together. In the next section I describe common themes in such systems, setting the stage for a theoretical framework relevant to this study of early implementation of Maine Performance Evaluation and Professional Growth systems.

Multi-Function Systems for Evaluation and Supervision

In the past decade or so, prepackaged programs—books, software systems, and other resources—have tried to combine evaluation and supervision into one cohesive system (Danielson, 2013; Marshall, 2013; Marzano & Toth, 2013). In brief, these models attempt to address both evaluation and supervision through cross-cutting use of numerous performance standards that address many facets of a teacher's work, detailed rubrics or scales that differentiate between levels of performance on those standards, professional development regarding the performance standards, observation tools and protocols constructed with the same standards, and often with data collection and management products available for purchase. Maine's Performance Evaluation and Professional Growth Systems regulation (Chapter 180,

2014, revised 2015) makes a similar combination and gives several of these prepackaged programs credence by pre-approving use to fulfill key regulatory functions.

Whether through anticipated efficacy, ease of adoption, or another factor, Maine school districts have followed in turn. In a 2015 "Intent to Pilot" survey 74% of Maine school districts that had reached consensus on professional practice standards for teachers indicated they had chosen either the Marzano (43%), Danielson (16%), or Marshall model (15%) for teachers (Lomonte, 2015). Thus, understanding characteristics of such multi-function systems is important in order to understand early implementation of Maine's mandates regarding teacher supervision and evaluation.

Systems intended to simultaneously address two complex and distinct functions are inherently complex. In this section I describe common characteristics of these multi-function systems intended to address teacher performance evaluation and teacher professional growth. The synthesis I present here does not include approaches that represent a minority of the literature, such as Peer Assistance and Review (PAR) systems. In this section I begin with broad aspects of system design and governance, then address more detailed elements in the order they may typically occur in an evaluation and supervision cycle.

Core Characteristics

Fundamentally, multi-function systems serve accountability purposes (Weisberg et al., 2009), with a focus on improvement of teachers, schools, and the school district (Bradshaw, 2002; Darling-Hammond, 2013; Marzano & Toth, 2013; Stronge & Tucker, 1999). In order to navigate the previously described tension between system rigidity and protection from arbitrary action, care is given to developing a system that specifies what is needed without specifying so many details as to detract from the overall goals (Darling-Hammond, 2013; DeSander, 2000;

Glickman et al., 2013; Koski, 2012). The system exhibits at least a strongly perceived level of validity: the system focuses on criteria important to effective performance, includes observable behavior that can be learned or changed, and excludes extraneous criteria not related to student performance or other goals of the organization (Bradshaw, 2002; Glickman et al., 2013; Holtzapple, 2003; Wilkerson et al, 2000).

Basic conditions of such systems include compliance with statutes, regulations, and collective bargaining agreements (DeSander, 2000) and a structure that addresses due process safeguards (DeSander, 2000; Platt et al., 2000; Sullivan & Zirkel, 1998). Further, the system is designed to be feasible with adequate resource support (Darling-Hammond, 2013; Marshall, 2013; Stronge & Tucker, 1999). Responsibility for the system is shared through a clear governance structure that includes teachers and administrators (Darling-Hammond, 2013; Marshall, 2013; Stronge & Tucker, 1999). Finally, the system design, governance, and implementation are conducted in such a way as to promote institutional conviction regarding student learning and effective teaching (Darling-Hammond, 2013; Platt et al., 2000; Stronge & Tucker, 1999; Weisberg et al., 2009), building trust, support, and collegiality (Glickman et al., 2013; Marshall, 2013; Murnane & Cohen, 1986; Stronge & Tucker, 1999).

Performance Standards, Rubrics, and Goals

The core of the multi-function system of supervision and evaluation is a carefully chosen set of performance standards; the standards provide the basis for all other components of evaluation and supervision (e.g., training and development of teachers, evaluators, and supervisors; classroom observations; summative evaluations). Without standards, measurement of teacher performance would be arbitrary (DeSander, 2000). As Platt et al. (2000) point out, "Administrators have wide-ranging definitions of what constitutes acceptable and excellent

teaching" (p. 32). Through these performance standards a school district expresses a shared and comprehensive model of effective teaching (Darling-Hammond, 2013; Marshall, 2013; Marzano & Toth, 2013; Medley & Coker, 1987; Platt et al., 2000). The standards communicate expectations in teachers' multiple professional roles, e.g., instruction, assessment, professional learning, and non-teaching responsibilities (Beckham, 1997; Darling-Hammond, 2013; DeSander, 2000; Platt et al., 2000; Stronge & Tucker, 1999). Ideally, the standards are also aligned with pre-service teacher standards and connected to evaluation systems for school and district leaders (Marzano & Toth, 2013; Weisberg et al., 2009), a relationship Marzano and Toth (2013) call "hierarchical evaluation" (p. 139).

While Hanushek and Rivkin (2006) note that researchers do not always agree on the characteristics of good teaching, the standards should be grounded in teacher effectiveness research (Darling-Hammond, 2013). Authors generally recommend weighting standards equally and using the same standards for novice and veteran teachers (Marshall, 2013; Marzano & Toth, 2013), though the expected level of performance on those standards may vary depending on teacher experience. Marzano and Toth (2013) express that using different criteria for different levels of experience could address the tendency noted above to over-rate novice teachers; Weisberg et al. (2009) would likely welcome this technique, sharing that overuse of high ratings makes high ratings meaningless. Additional or alternative standards may need to be developed to address professional specializations (Spillane et al., 2002).

The teacher performance standards are further detailed into rubrics or scales, describing explicitly what it means to satisfy that standard (Darling-Hammond, 2013; Marshall, 2013). In contrast to binary (satisfactory/unsatisfactory) ratings, these rubrics typically use four to five levels of performance to provide a continuum of ratings options (Darling-Hammond, 2013;

Holtzapple, 2003; Marshall, 2013; Marzano & Toth, 2013; Medley & Coker, 1987; Weisberg et al., 2009), shown in this example:

The teacher:	4 Highly Effective	3 Effective	2 Improvement Necessary	1 Does Not Meet Standards
a. Expectations	Is direct, specific, consistent, and tenacious in communicating and enforcing very high expectations.	Clearly communicates and consistently enforces high standards for student behavior.	Announces and posts classroom rules and consequences.	Comes up with *ad hoc* rules and consequences as events unfold during the year.
b. Relationships	Shows warmth, caring, respect, and fairness for all students and builds strong relationships.	Is fair and respectful toward students and builds positive relationships.	Is fair and respectful toward most students and builds positive relationships with some.	Is sometimes harsh, unfair, and disrespectful with students and/or plays favorites.

Figure 2.2. Example rubric elements for Domain B: Classroom Management. These two examples from Marshall (2014) are a subset of the 10 rubric elements in this domain.

For the professional growth function the rubrics provide a map for improvement; whether below standard, mediocre, or effective, a teacher and his or her supervisor can look to the next level's description of performance to see what they must do to develop skill in that area. For the performance evaluation function rubrics inform and provide efficiency to documentation; selecting rubric ratings eliminates some need for lengthy narratives (Marshall, 2013).

Within the context of the school district's goals (expressed in performance standards), the individual teacher's goals play a large role in the multi-function supervision and evaluation system. Sources of information that individuals use in defining their personal goals include the previous summative evaluation and a self-audit or self-assessment (Marshall, 2013; Marzano & Toth, 2013; Platt et al., 2000). Methods used in this process include, e.g., videotaping classrooms (Glickman et al., 2013), portfolio development (Zepeda, 2002), and comparing practices by visiting peers' classrooms (Marzano & Toth, 2013). Typically the teacher proposes goals, which are then accepted or modified by the evaluator and/or supervisor (Darling-Hammond, 2013; Marzano & Toth, 2013). Through this individual focus, the system promotes buy-in and reflection (Marshall, 2013; Marzano & Toth, 2013; Platt et al., 2000).

Ongoing Collection of Evidence and Instructional Feedback

With standards and rubrics in place to define and describe the school district's model of effective teaching, the multi-function system of supervision and evaluation must then attend to the types of data and evidence that teachers, evaluators, and supervisors will use in the process. The traditional system focused on only one or a few classroom observations (Wilkerson et al., 2000); authors now call for multiple sources of data (Darling-Hammond, 2013; Marzano & Toth, 2013; Platt et al., 2000). As Wilkerson et al. (2000) point out, "Teaching is such a complex and contextualized phenomenon that any single mode of measurement [fails] to assess true teacher performance" (p. 180). Various authors lean toward different sources of evidence and generally opt to include rather than prohibit possible data sources, including the examples shown in Table 2.2 below.

Several sources of evidence appear repeatedly in the literature: observations of classroom or non-classroom activities by the evaluator or peers (Marshall, 2013; Marzano & Toth, 2013; Platt et al., 2000), plans and analysis for units of instruction (Darling-Hammond, 2013; Marzano & Toth, 2013; Platt et al., 2000), teacher portfolios (Darling-Hammond, 2013; DeSander, 2000; Glickman et al., 2013; Platt et al., 2000; Stronge & Tucker, 1999), and evidence of professional learning (Platt et al., 2000; Wilkerson et al., 2000). Authors generally call for some inclusion of direct evidence of student learning in teacher supervision and evaluation systems (Platt et al., 2000), especially focusing on teacher analysis of student achievement data (Darling-Hammond, 2013; Glickman et al., 2013; Marzano & Toth, 2013). Whether and how to include student assessment scores as a direct factor in teacher evaluation is a matter of some debate in the literature.

Table 2.2. *Examples of Supervision and Evaluation Data Sources*

Teacher…	Classroom Artifacts Such As…
Analysis of an Instructional Unit	Assessments (Formative or Summative)
Arrival/Departure Times	Assignments
Attendance Profiles	Discipline Referrals
Co-Curricular Participation	Gradebooks
Content Knowledge	Grading Criteria
Performance on Standardized Assessments	Handouts
Portfolios	Newsletters or Memos
Professional Development	Progress Report Comments
	Unit or Lesson Plans
Student…	Observations, Scales, or Surveys From…
Grade Distributions	Department Chair or Team Leader
Growth Over Academic Year	Other Administrators
Interviews	Parents
Performance on Class Assignments	Peers/Colleagues
Performance on Standardized Assessments	Principal or Assistant Principal
Placement Requests	Self
Work Samples with Teacher Feedback	Students

Note. These examples of evidence and data sources for teacher supervision and evaluation are drawn from a variety of authors (Darling-Hammond, 2013; DeSander, 2000; Glickman et al., 2013; Marshall, 2013; Marzano & Toth, 2013; Platt et al., 2000; Stronge & Ostrander, 1997; Stronge & Tucker, 1999; Wilkerson et al., 2000)

Gathering evidence through classroom observations receives significant attention in the literature as the primary data source for teacher supervision and evaluation. The observation system is deliberately designed to reduce sampling error (Marzano & Toth, 2013). The system may require a certain number of observations for evaluative purposes; when whole-class observations are used systems typically require one to three annual observations (Marzano & Toth, 2013; Platt et al., 2000). However, "one high-stakes observation a year has a high probability of getting an inaccurate picture of daily reality and raising the teacher's anxiety level to stratospheric heights" (Marshall, 2013, p. 58); sampling error is high with four or fewer observations (Marzano & Toth, 2013). Authors in the literature generally prefer multiple

observations; a greater number of observations increases accuracy and decreases sampling error (Darling-Hammond, 2013; Glickman et al., 2013; Marshall, 2013; Marzano & Toth, 2013; Platt et al., 2000). More frequent observations may also lead to greater teacher acceptance of the process (Oliva & Pawlas, 2004). Marshall (2013) notes a tension between the frequency and duration of observations, recommending approximately ten "mini-observations" annually (each ten or more minutes in length with prompt feedback). With this choice, whole-class observations are reserved for certain instances such as novice teachers or teachers at risk of dismissal. Synthesizing the recommended models in the literature results in a mixture of whole-class and partial-class observations, some announced in advance and some unannounced observations for which the teacher does not have the opportunity to specially prepare (Marshall, 2013; Marzano & Toth, 2013; Platt et al., 2000).

In addition to multiple observations, the observation system may include multiple observers to reduce measurement error (Marzano & Toth, 2013). These additional observers may include individuals such as other administrators, department/team leaders, or peers. If this is done, authors recommend focusing on thorough training and protocols; peer observations in particular pose challenges with leniency (Marshall, 2013; Platt et al., 2000).

Whether one or multiple observers are used in the teacher supervision and evaluation system, protocols and instruments (based on the performance standards and rubrics described above) guide the process. Many possible observation instruments are discussed in the literature. Glickman, Gordon, and Ross-Gordon (2013), e.g., identify the following instrument possibilities for varied purposes: visual diagramming, open-ended, verbatim, selected verbatim, forced questionnaire, performance indicator, and categorical frequency. Generally, however, the instruments required in the system should be less limiting than a checklist approach (Darling-

Hammond, 2013) and should provide sufficient detail to differentiate between skill levels (Holtzapple, 2003; Marzano & Toth, 2013). When selecting or developing observation instruments, care should be given to expect professional judgment from teacher and observer rather than attempt to limit judgment through checklists; some lists may lead to use of canned routines whether or not they are appropriate to the instructional circumstance (Darling-Hammond, 2013).

Protocols developed in the teacher supervision and evaluation system may include pre- and post-observation conferences and may include mutual viewing of videotape by the teacher and observer (Darling-Hammond, 2013; Marzano & Toth, 2013; Platt et al., 2000). Formative feedback is essential; face-to-face discussion should occur before the observation is documented (Marshall, 2013; Platt et al., 2000). Darling-Hammond (2013, p. 53) writes, "Research has found that the frequent, skilled use of standards-based observation with feedback to the teacher is significantly related to student achievement gains, as the process helps teachers improve their practice and effectiveness." Other factors that can inform the observer include learning how the observed lesson fits into the unit and the year's instructional goals (Marshall, 2013), asking students in the class about what they are learning (Platt et al., 2000), and observing out-of-class situations (e.g., duties, parent conferences, meetings) to gather evidence of contributions to the school as a whole (Darling-Hammond, 2013; Platt et al., 2000).

Documentation of Evidence-Gathering

Documentation of observations for formative or summative purposes can take several forms; generally, the documentation expectations should be detailed in the system, and authors recommend separating description from interpretation.

> Sharing the description of events is the forerunner of professional improvement. Interpretation leads to resistance. When both parties can agree on what events occurred, they are more likely to agree on what needs to be changed. (Glickman et al., 2010, p. 237)

Factors documented may include data gathering procedures, the perceived instructional impact, commendations or reinforcement of sought teaching behaviors, suggestions or recommendations to improve instructional practice, direction when a change in practice is required, and an overall rating (Glickman et al., 2013; Marshall, 2013; Platt et al., 2000). Technology platforms may be used during an observation to aid in efficiency (Danielson, 2013; Marzano & Toth, 2013) but are not universally recommended (Marshall, 2013).

Through all phases of evidence-gathering (classroom observations and/or other data) some ongoing documentation is important for improvement and assessment purposes (Beckham, 1997; DeSander, 2000). Authors (Darling-Hammond, 2013; Marshall, 2013) recommend designing systems that minimize paperwork for all involved. For formative improvement purposes, documentation that keeps track of individual progress (Marzano & Toth, 2013) helps to show teachers where they are going and where they need to grow. Those same data, compiled at a school or district level, can help to set school or multi-school goals and identify professional development needs (Marshall, 2013).

For summative evaluation purposes that potentially include negative outcomes such as discipline or dismissal, documentation is particularly important (Sullivan & Zirkel, 1998). On any occasion where a rating is less than satisfactory, documentation of evidence to support that rating and establish a pattern is critical:

> The prevailing view is… behavior that wasn't serious enough for the supervisor to have documented when it happened is not important enough to be considered later as evidence to support disciplinary action. Generally speaking, in an employee discipline case, the attempt to prove a fact that was not documented and shared with the employee involved is likely to be viewed as an after-the-fact effort to justify a subjective decision that has already been made. (Platt et al., 2000, p. 54)

For less than satisfactory conditions, the documentation should include evidence, an explicit problem statement, and a timetable for the next step (Platt et al., 2000). For any teacher, high-performing or otherwise, the focus in the literature remains on designing systems that provide useful information for continual improvement of professional performance.

Ongoing Training and Development

Echoing the previously stated emphasis on improvement (while still serving the accountability function), multi-function systems of teacher supervision and evaluation include a large focus on training and professional development. For teachers, the recommended initial training focuses on developing understanding of the system, especially on the specific performance criteria (Beckham, 1997; DeSander, 2000; Platt et al., 2000). Ongoing teacher development is recommended to directly and immediately connect to the classroom observation through feedback, use of the clinical supervision model of pre- and post-observation conferences, or through mutual viewing of videotape by the teacher and observer (Darling-Hammond, 2013; Marzano & Toth, 2013; Platt et al., 2000).

For supervisors and evaluators, the multi-function supervision and evaluation system requires more extensive training (Bradshaw, 2002; Darling-Hammond, 2013; DeSander, 2000; Marshall, 2013; Medley & Coker, 1987; Weisberg et al., 2009), conducted through a systematic

approach (Marzano & Toth, 2013). Multiple authors repeatedly emphasize the critical nature of training; Platt et al. (2000), for example, recommend that districts "invest significant time and energy developing professional community and technical competence among supervisors and evaluators" (p. 30). Weisberg et al. (2009) continue: "Administrators must receive rigorous training and ongoing support so that they can make fair and consistent assessments of performance against established standards and provide constructive feedback and differentiated support to teachers" (p. 7). Regarding the assessment of teacher performance, authors focus on training to reduce leniency or severity measurement errors (Platt et al., 2000) thus increasing inter-rater reliability (Bradshaw, 2002; Darling-Hammond, 2013; DeSander, 2000; Glickman et al., 2013; Marshall, 2013; Marzano & Toth, 2013; Platt et al., 2000). Methods used in this training may include a focus on type and level of teaching strategies (Marzano & Toth, 2013), group efforts such as visiting teams or grand rounds (Platt et al., 2000), and collective viewing of classroom video followed by discussion (Darling-Hammond, 2013; Marshall, 2013; Marzano & Toth, 2013).

Specific supervisor and evaluator training extends beyond the classroom observation and measurement of performance standards. Training to conduct post-conference sessions (Darling-Hammond, 2013) prepares supervisors and evaluators to give effective feedback, and critiquing others' evaluation reports (Platt et al., 2000) enhances skill in written documentation. Finally, the system synthesized from recommendations in the literature does not assume ongoing proficiency, but maintains accountability through a technique such as annual recertification (Darling-Hammond, 2013) or regular monitoring of evaluator judgments (Weisberg et al., 2009).

Summative Evaluation, Consequences, and Goal-Setting

The summative evaluation synthesizes the evidence gathered throughout the year (Platt et al., 2000). In contrast to binary ratings (e.g., satisfactory–unsatisfactory), multiple rating steps more accurately differentiate teacher effectiveness (Marzano & Toth, 2013), provide motivation for teachers to enhance practice, and leave room for documented improvement (Marshall, 2013). Marshall, for example, notes that "binary ratings are unlikely to motivate a mediocre teacher to take it up a notch or to spur a good teacher to strive for excellence" (p. 24), and recommends a rating system with four levels: (1) Does Not Meet Standards, (2) Improvement Necessary, (3) Effective, and (4) Highly Effective (p. 129). The system, however, should recognize that time and practice are needed to develop expertise: "A teacher evaluation system focused on development would not expect teachers to reach the highest levels of effectiveness quickly" (Marzano & Toth, 2013, p. 103). Marzano and Toth raise the possibility of different summative criteria for different levels of experience, though others (Darling-Hammond, 2013) recommend using consistent standards for different phases of training and development.

While there is not a single approach preferred in the literature, the system should specify how the identified sources of evidence are combined into a summative rating. Marzano and Toth (2013), for example, discuss compensatory versus conjunctive approaches. In a compensatory approach high scores on some factors may compensate for low scores on other factors, while in a conjunctive approach threshold or cut scores emphasize the relative importance of individual factors (Marzano & Toth, 2013). Generally, care must be taken to avoid an approach in which strengths can mask significant weaknesses (Marzano & Toth, 2013; Platt et al., 2000).

A summative evaluation meeting between the teacher and evaluator for face-to-face discussion includes comparison of a teacher's self-assessment to the evaluator's assessment

(Marzano & Toth, 2013; Platt et al., 2000); this may be done simultaneously through a mutual "reveal" (Marshall, 2013, p. 152) to avoid perception that the evaluator changed ratings based on the teacher's self-assessment. Documentation of the summative evaluation includes goals, data gathering procedures, judgments, comments, an overall rating, and a recommendation regarding personnel action (Platt et al., 2000). Strengths and problems should be specified; if deficiencies are noted in the summative evaluation they should be clearly communicated and backed with evidence (Platt et al., 2000; Sullivan & Zirkel, 1998).

The summative evaluation conference (identifying strengths and needs) may then lead directly to goal setting for the next supervision and evaluation cycle and/or to opportunities or consequences. Examples of areas that could be formally tied to evaluation ratings include compensation changes, a varied supervision model for the next evaluation cycle, identification for intensive assistance, identification for mentoring of others, layoff order, and retention or contract decisions (Darling-Hammond, 2013; DeSander, 2000; Marzano & Toth, 2013; Murnane & Cohen, 1986; Weisberg et al., 2009). While other examples are noted in the literature, authors focus on outcomes for high-achievers and for those performing below expectations.

For teachers performing highly relative to the shared definition of excellence (Platt et al., 2000), authors generally promote recognition (Darling-Hammond, 2013; Weisberg et al., 2009) and opportunities for advancement, leadership, or to spread expertise (Darling-Hammond, 2013; Marshall, 2013). Merit pay, however, receives considerable debate in the literature and in the public arena. Merit pay is "any system of teachers' compensation that explicitly rewards better performance" (Dee & Keys, 2004, p. 473), potentially delivered through individual bonuses or group incentives (DeSander, 2000). While deep discussion of merit pay is beyond the scope of this paper, and public opinion and rhetoric may differ, it is important to briefly note here that

multiple authors have written about significant challenges in the use of merit pay systems in the teaching profession (Dee & Keys, 2004; DeSander, 2000; Eberts, 2007; Koski, 2012; Murnane & Cohen, 1986).

The literature is relatively cohesive regarding steps following an unsatisfactory summative evaluation: clear communication, intensive support, ongoing documentation, and regular formal re-evaluation to determine employment outcomes. "Providing struggling teachers with a detailed diagnosis and prescription and a chance to improve is an important ethical and legal responsibility," says Marshall (2013, p. 149) but the system must "fairly but swiftly remove consistently low-performing teachers" (Weisberg et al., 2009, p. 30; echoed by Darling-Hammond, 2013; Marzano & Toth, 2013; Platt et al., 2000). If incompetence, deficiency, or mediocrity is identified in the summative evaluation, the supervision and evaluation system synthesized from recommendations in the literature first calls for the evaluator to clearly and explicitly state the problem or problems (DeSander, 2000; Glickman et al., 2013; Platt et al., 2000). While all problems should be listed to avoid later claims of harassment, the problems may be ranked or prioritized (Platt et al., 2000). In this documentation, the evaluator must be clear about expectations versus suggestions, and "mediocrity is protected by poor or tentative writing and mixed messages" (Platt et al., 2000, p. 25). The teacher is given a formal opportunity to respond to the notice of problems (Glickman et al., 2013).

Following the notice of problems, attention immediately turns to developing a written improvement plan to address the identified needs (Beckham, 1997; Glickman et al., 2013; Platt et al., 2000). Courts tend to enforce requirements for credible improvement or remediation plans (Sullivan & Zirkel, 1998). The plan should specifically focus on the performance standards of the system, and on performance outcomes rather than professional development activities. Platt

et al. (2000) say, “Performance improvement goals are statements of what needs to happen to eliminate the problem and improve learning for students” (p. 129).

Timelines in improvement plans vary based on the problems identified and must be reasonable (Beckham, 1997; DeSander, 2000; Medley & Coker, 1987; Platt et al., 2000). Marshall (2013) recommends that initial interventions be identified for a short timeline such as one month, while the full plan timeline may span a year (Sullivan & Zirkel, 1998). Platt et al. (2000) provide a legal observation that the longer teachers have been employed by a district, the more time they must generally be given to remediate deficiencies.

When possible, the teacher and evaluator should mutually develop the improvement plan (Platt et al., 2000). Support for the teacher should be identified and documented, possibly including supervisors, an assistance team, or peer mentors (Darling-Hammond, 2013; Glickman et al., 2013; Platt et al., 2000). The improvement plan should also include the data sources that will be used as evidence of progress and the timeline for periodic reevaluation (Glickman et al., 2013; Platt et al., 2000). Additional evaluators (and/or supervisors) may be brought in to provide more data and support (Beckham, 1997; Platt et al., 2000). After improvement is documented, follow-through should continue to prevent backsliding (Platt et al., 2000).

Prior to dismissal, systems may call for disciplinary steps such as salary freezes (Platt et al., 2000; Weisberg et al., 2009). In the event that the teacher does not improve and dismissal is sought as an employment outcome, the system should anticipate arbitration and include documentation that establishes patterns over time, progressive discipline, and overall processes that satisfy the *just cause* standard required by many states or collective bargaining agreements (Marshall, 2013; Platt et al., 2000; Weisberg et al., 2009). Careful attention to every step is

necessary; “procedural errors still result in the reinstatement of teachers” (Sullivan & Zirkel, 1998, p. 379).

Resources Needed for Teacher Supervision and Evaluation Systems

As described above, the multi-function teacher supervision and evaluation system synthesized from recommendations in the literature includes many components and requires much of school districts, supervisors, evaluators, and teachers. Developing these systems requires significant time and human resources (Bradshaw, 2002; Darling-Hammond, 2013); the same and additional resources are critical to implement and sustain systems. For teachers, if the system is to connect powerfully to professional learning the school district must dedicate supportive resources such as curriculum specialists, supervisors, funds for courses or workshops, release time for professional learning, and peer assistance (Stronge & Tucker, 1999). Time to collaborate, plan, and learn with colleagues is also recommended; Darling-Hammond (2013) notes that teachers in the United States average far fewer weekly hours for collaboration relative to other developed nations.

For supervisors and evaluators, implementation of the system requires considerable time (Darling-Hammond, 2013; Marshall, 2013). The school district must give attention to manageable caseloads (Marzano & Toth, 2013) through hiring sufficient human resources for implementation. In addition to the typical supervisors and evaluators (e.g., instructional supervisors, principals), systems may draw in assistant principals, department heads, mentor teachers, and consulting teachers to address the time challenge and provide additional support (Darling-Hammond, 2013; Goldstein, 2007). School districts can additionally assist supervisors and evaluators with time management (Marshall, 2013) as they take on roles not typically previously required. If technological platforms are to be part of the teacher supervision and

evaluation system (Danielson, 2013; Marzano & Toth, 2013), funds must be dedicated to purchase and maintain these systems. Whether through computer platforms or through low-tech means, time resources must be dedicated to charting faculty results to identify school and district strengths and needs (Marshall, 2013).

To sustain multi-function supervision and evaluation systems, authors note that school districts must continue to dedicate resources; American school history is replete with examples of unsustained efforts (Tyack & Cuban, 1995). The time and funding resources dedicated to initial training of teachers, evaluators, and supervisors must continue (Bradshaw, 2002) in order to bring new staff into the system, to reinforce expectations with existing staff, and to institutionalize the system over time (Marzano & Toth, 2013). Resources must be dedicated to monitor the system (Bradshaw, 2002; Weisberg et al., 2009), holding all accountable for their responsibilities and ensuring the agreed-upon professional standards are maintained. Furthermore, school districts must dedicate resources to analyzing the system results and ongoing refinement of the system (Marzano & Toth, 2013; Stronge, 1997).

Bradshaw (2002) studied 27 North Carolina school districts in the wake of a statewide evaluation law, and concluded:

> Without ongoing monitoring and evaluation and with decreasing resources to support the teacher evaluation system, discrepancies in the use of the NC TPAS grew within and across districts. Ratings became inflated, and the administrators conducting teacher evaluation were less likely to be sufficiently trained. Requirements for classroom observations and other activities were ignored. (p. 124)

While Bradshaw focused above on declining state resources for monitoring and evaluation of statewide teacher evaluation systems, declines in dedicated district resources to support, monitor, and refine teacher supervision and evaluation systems may have similar detrimental effects.

Professional Growth Focus in Multi-Function Supervision and Evaluation Systems

Weisberg et al. (2009) write about important connections between evaluation and supervision functions: the inability to accurately assess instructional practice and act on that information prevents schools from providing differentiated professional development to ineffective teachers and "supporting growth among the broad plurality of hard-working teachers who operate in the middle of the performance spectrum" (p. 2). Summative evaluation supplanting supervision is problematic, however, summative and formative components of a system can work in concert with the agreed-upon standards for excellent teaching to provide the teacher with valuable feedback and direction for professional growth (Holtzapple, 2003). While the sample was not scientifically representative, Marzano and Toth (2013) found that 76% of respondents valued teacher development over measurement in classroom observations. Oliva and Pawlas (2004) share a similar conclusion: "We find a definite acceptance of the idea…to help teachers build on their strengths, improve, and remain in the profession instead of probing teachers' deficiencies and seeking their dismissal" (p. 9). Zepeda (2006) also speaks to supervisory needs: "Teachers need support and leaders willing to make supervision a precursor to annual evaluation. The intents behind supervision and evaluation are quite different; however, evaluation without supervision first smacks of professional malpractice" (p. 68).

In designing a multi-function system where supervision and evaluation work in concert toward professional growth, a focus should remain on assessing and providing valuable feedback with regards to what happens in the classroom as a means for positively influencing practice:

"observations should be used as a base of information to create an instructional dialogue between supervisor and teacher" (Glickman et al., 2010, p. 237). Effective feedback to teachers is frequent (Marshall, 2013; Weisberg et al., 2009), specific (Marzano & Toth, 2013; Platt et al., 2000), and attuned to the needs of adult learners (Ponticell & Zepeda, 2004). Weisberg et al. (2009) represent the literature well: "Constructive feedback that specifies areas for development is a critical facet of any performance evaluation, even for strong performers" (p. 14). Feedback, however, must not only be provided annually for summative evaluation purposes, but also for the supervision functions of professional development and support:

> Evaluation alone will not improve practice. Productive feedback must be accompanied by opportunities to learn. Evaluations should trigger continuous goal-setting for areas teachers want to work on, specific professional development supports and coaching, and opportunities to share expertise, as part of recognizing teachers' strengths and needs…evaluation can be used to stimulate meaningful professional learning as teachers set goals and pursue them with the assistance of administrators and colleagues. In addition, it can be used to flag areas for further support that are made available through a cycle of ongoing professional development. (Darling-Hammond, 2013, p. 99)

Zepeda (2012) lays out a growth-focused model similar to Darling-Hammond's assertion, shown in Figure 2.3 below, with professional development at the center of activities including supervision and evaluation.

The approaches by Zepeda and Darling-Hammond are supported by others in the literature. Authors recommend professional development highly focused on individual needs and goals (Marzano & Toth, 2013; Weisberg et al., 2009), influenced and incentivized by evaluative feedback (Darling-Hammond, 2013; Holtzapple, 2003). Professional development

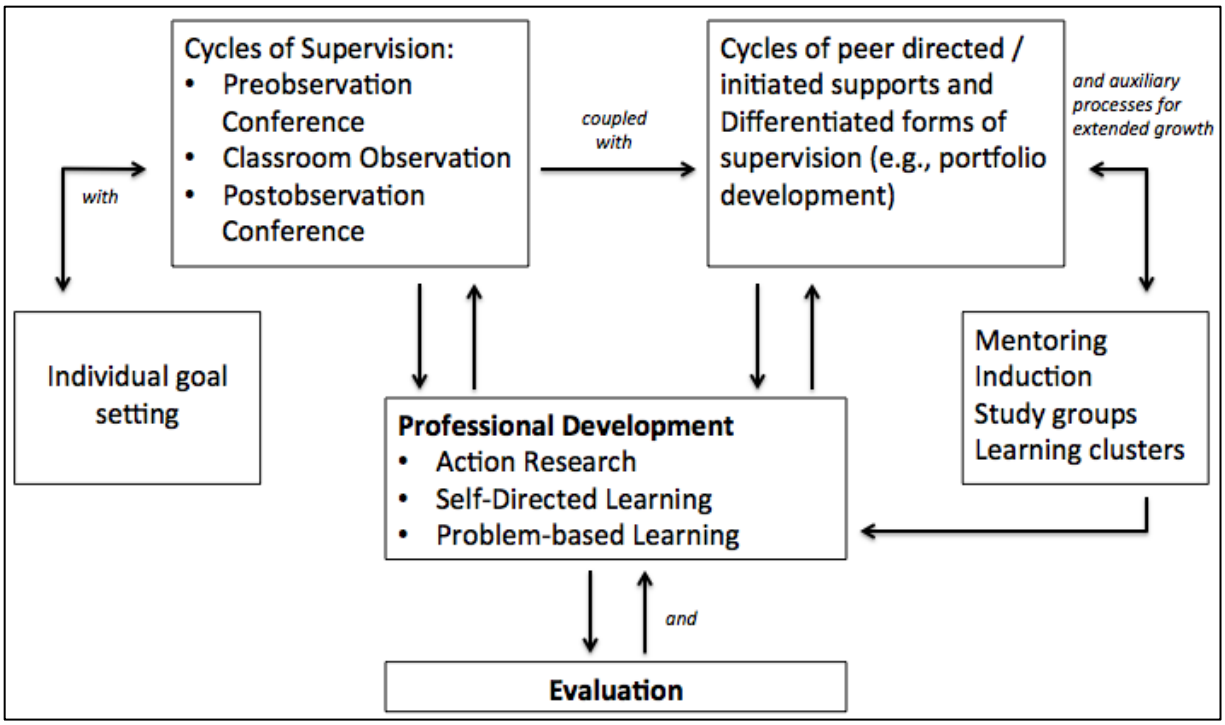

Figure 2.3. One unified model for teacher growth (Zepeda, 2012, p. 18).
In this model Zepeda links supervision, evaluation, and professional development.

opportunities should be linked to the agreed-upon performance standards (Darling-Hammond, 2013; Weisberg et al., 2009), with significant time and support available (Darling-Hammond, 2013). Trained peer coaches or mentor teachers can be particularly effective sources of guidance and expertise (Darling-Hammond, 2013; Glickman et al., 2013; Marshall, 2013; Marzano & Toth, 2013). Similarly, effective professional development systems coordinated with evaluation systems may focus on collaborative, team-based strategies (Darling-Hammond, 2013; Marshall, 2013; Marzano & Toth, 2013; Platt et al., 2000). "In complex endeavors such as teaching," Marzano and Toth (2013) said, "it is extremely difficult to reach and then maintain the highest levels of performance without help" (p. 128). Opportunities for differentiated, job-embedded learning connected to reflection, collaboration, and dialogue increase the potential for professional development to positively impact teachers and students (Zepeda, 2006).

As a whole, while some authors view supervision and evaluation as distinct and mutually exclusive, others posit that a system can be created that fulfills the personnel function (i.e., evaluation) but also focuses on goals and growth for the development of all teachers. In brief, these systems use a core of comprehensive performance standards and rubrics in conjunction with organizational and individual goal-setting, collection of evidence, instructional feedback, training and development, and summative evaluation. In contrast to several authors above, Glickman, Gordon, and Ross-Gordon (2013) do not sell a pre-packed system of teacher supervision and evaluation but promote a vision—"SuperVision" in their words—for adult interactions resulting in successful collegial schools.

"SuperVision" for Successful Schools

Glickman, Gordon, and Ross-Gordon (2013) provide a theoretical lens by which to understand systems of evaluation and supervision, positing that schools must transition to a *collegial* model for success. Glickman et al. (2013) begin with description of a history of *conventional* (p. 7) school adult interactions, focused on inspection and attempts to control teacher behaviors. In their illustrative example, the school principal details materials, schedules, and timelines, enforcing instructions and commonality with little tolerance for variation from the plan. Glickman et al. outline limitations to these control-oriented or directive-oriented approaches: "dependency, hierarchy, and professional isolation" (p. 6) ultimately resulting in the departure of teachers from the profession. They do not, however, promote the opposite approach, also describing the limitations of *congenial* schools in which there is little direction or commonality, with a focus on "friendly social interactions and professional isolation" (p. 6). In such congenial schools, individual teachers "seem to have the discretion to function as they

please" (p. 4), which may mean long-term use of ineffective instructional practices or inefficient use of time.

In the theoretical view of Glickman et al. (2013), successful schools utilize a *collegial* approach "characterized by purposeful adult interactions about improving schoolwide teaching and learning" (p. 6). While they describe essential elements for summative evaluation and recognize its function, the researchers keep a focus on collaboratively generating and implementing a school vision for teaching and learning and providing quality instructional supervision: "assistance for the enhancement of teaching and learning" (p. 9). They describe a model in which democratic decision-making flourishes and hierarchy fades, and in which individuals and schools engage in continual self-study to implement and refine the shared vision. Successful supervision and evaluation, then, relies on individual and school goals working "in harmony toward [the] vision of what the school *should* be" (2013, p. 9, emphasis in the original).

Indicators of this "SuperVision" model, synthesized from Glickman et al. (2013), are as follows:

- A collaboratively developed vision for teaching and learning
- Collaborative implementation of the vision for teaching and learning
- Supervision that includes teachers and formal supervisors, in a way that minimizes hierarchy and maximizes collegiality
- "A focus on teacher growth rather than teacher compliance" (p. 7)
- Facilitation of teacher collaboration in instructional improvement and reflective inquiry
- A focus on "communal leadership" rather than "heroic individuals" (p. 10)
- Deliberate development of knowledge, interpersonal skills, and technical skills to support these efforts.

Glickman et al. (2013) lay out this model, from prerequisites to the student-focused outcome, as follows:

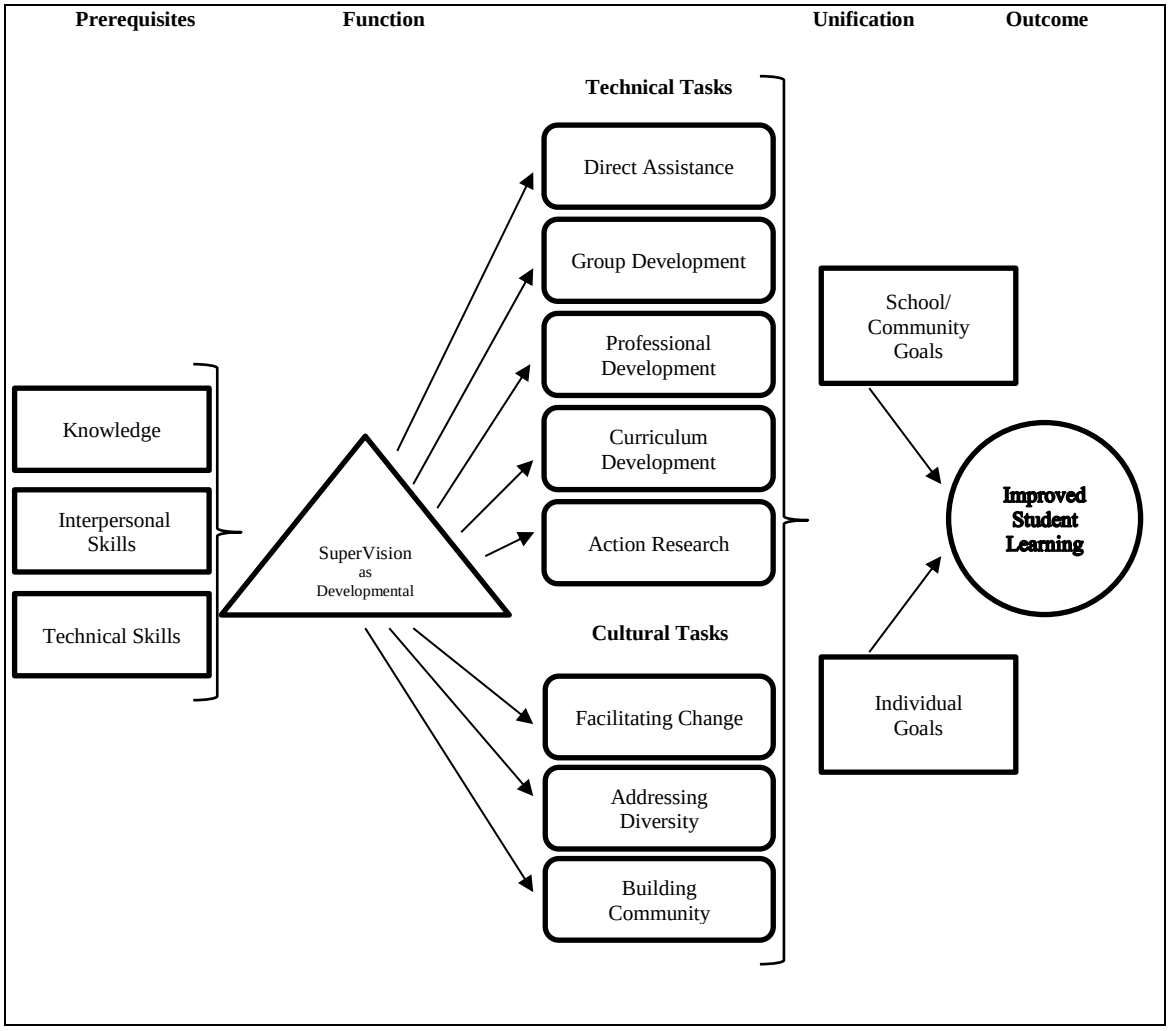

Figure 2.4. SuperVision (Glickman et al., 2013, p. 14)

Glickman et al.'s vision (2013) generally shares outcome goals with Maine's PE & PG systems regulation, though the supervision-focused approach is largely in contrast to Chapter 180's text focus on summative evaluation. Glickman et al. (2010) caution that combined summative evaluation and formative evaluation (i.e., supervision) systems tend to inadvertently focus on the summative evaluation function: [S]ummative rating scales are designed to be standardized, global, legally defensible, efficiently completed and processed….this means not only that the ratings have little value for [supervision] but also that the richest, most meaningful data for

[supervision] is precluded" (p. 277). *SuperVision* is generally representative of the theoretical literature's focus on supervision as the primary means by which to improve teaching and learning, and is a major component of the theoretical framework for this study. That framework will be presented below after a brief review of the national and state research that has been done so far on similar topics.

Emerging Research in the Era of Evaluation and Supervision Mandates

The 2009 *Race to the Top* program and 2011 *No Child Left Behind* Flexibility Waivers pushed many states to implement changes to systems of teacher evaluation or systems of teacher evaluation and supervision (Kraft & Gilmour, 2017). The details of these systems vary state-by-state, with common themes of professional standards, detailed rubrics, student growth data, training and calibration for observers, and overall systems much more comprehensive than the preceding systems. A small but growing body of literature explores the nature and impact of these systems; most of the research focuses on the perspectives of evaluators rather than teachers or supervisors. Below I briefly review single-state and multi-state research relevant to my study before focusing on the studies conducted by the Maine Education Policy Research Institute.

Ramirez, Clouse, and Davis conducted a 2014 study intended to identify teacher evaluation barriers in Colorado and make policy recommendations. The study utilized a mixed-methods design with focus group interviews of administrators, surveys of teachers who were graduate students in leadership programs (n = 108), surveys of human resources personnel (n = 46), and surveys of administrators (n = 402). Among other barriers, they found time requirements were the most major obstacle to teacher evaluation. Kowalski and Dolph (2015) reached similar conclusions, surveying 50 principals in Ohio in the context of the new Ohio Teacher Evaluation System. They found considerable concern with excessive time demands on

the principals (by 96% of respondents), excessive time demands on teachers (72% of respondents), and also considerable "opposition to using student value-added measures" (2015, p. 14).

However, the time demands may not be insurmountable. Derrington and Campbell (2015) conducted a three-year multi-site qualitative study in an unnamed state which received a *Race to the Top* grant that required changing teacher evaluation systems. Their study included 14 principals in four school districts (suburban and rural). In Year 1, all principals in the study described lack of time as their greatest frustration. There was some response by districts, assigning central office personnel to share the burden of the evaluation process. In the second year of the study half of the principals reported that the extensive time demands remained; fewer still reported extensive time concerns in the third year, having developed routines, procedures, and strategies to address the challenge. The system was viewed by the principals in the third year as "increasingly beneficial, largely with regard to the degree to which the rubric was helping to inform and improve instruction" (p. 317).

Goldring et al. (2015) also found that principals favored the use of common and detailed rubrics paired with teacher observations. Their study used surveys and semi-structured interviews to yield the perspectives of central office personnel and principals in six urban districts (multiple states) implementing new teacher evaluation and supervision systems from 2012–2014. The same study shed light on unintended consequences of the use of value-added measures (VAM) for student growth data. First, "[S]ome principals did use VAM to move ineffective teachers to untested grades, such as K-2 in elementary schools and 12th grade in high schools" (p. 100). Second, in half of the systems they studied, principals were held accountable for their observation scores correlating with the state test scores. A district leader shared, "[The

principals] know that their evaluation is dependent on the correlation, so they're just looking at the [test scores from] the year before…while they're making their observations. So total invalidation of the observation process" (p. 101). While noting that "many principals are simply overwhelmed" (p. 102) by time-consuming requirements, Goldring et al. found that the principals preferred observations with detailed rubrics over value-added measures for purposes of specific and actionable feedback, timeliness, utility across subject matter, and clarity.

However, others challenge the connection between observations and actual instructional effectiveness, even when using a measurement instrument such as a rubric. In the wake of *Race to the Top*, Strong, Gargani, and Hacifazlioglu (2011) conducted a series of three experiments in California, Tennessee, and Virginia to explore the relationship between teacher observation and known effectiveness (defined in this case through value-added measures). Their study included hundreds of expert and non-expert judges (e.g., school administrators, math teachers, adults with no school affiliation) viewing videotape of instruction ranging from short excerpts to full lessons. Overall they found "high agreement among judges but low ability to identify effective teachers" (p. 367), with results no more accurate than chance and little relationship between expertise and accuracy. They did note some sub-scores on a rubric that were potentially promising in the quest to develop an observational measure that could predict effectiveness, and also shared that standardized test scores were a narrow answer to the question, "How should we define effective teaching?" (p. 379).

Kraft and Gilmour (2017) were also open to such fundamental questions, remarking that for some the purpose of evaluation is primarily to dismiss low-performing teachers, and for others the purpose of evaluation is primarily growth through feedback and targeted support. They asserted that both functions require evaluation systems that accurately assess instructional

quality and differentiate among teachers. In a study involving 24 states, they sought to determine the extent to which new teacher evaluation systems differentiate among teachers, explicitly as a follow-up to The New Teacher Project's *The Widget Effect* (Weisberg et al., 2009). They found that overall the percentage of teachers rated unsatisfactory had not changed in most states using new teacher evaluation systems, but that there was great variation across states (e.g., those rated below proficient ranged from a low of 0.7% in Hawaii to 28.7% in New Mexico, those rated above proficient ranged from 6% in Georgia to 62% in Tennessee). They also found that principals tended to rate teachers at the summative evaluation more generously then the principals believed were truly warranted. Exploring this further, Kraft and Gilmour (2017) found "persistent implementation challenges" and "competing trade-offs" (p. 240), such as time constraints (especially as pertains to providing intensive support and documentation), belief in teachers' potential and motivation, personal discomfort, trading proficient evaluations for voluntary departure, and principals fearing lower-quality replacements if they dismissed a teacher.

Hallinger, Heck, and Murphy (2014) conducted a critical evaluation of the evidence on what they termed "new generation teacher evaluation systems" (p. 5), including standards-based measures of teacher quality and measures of student learning.[4] They found "a pattern of weak, inconsistent, and unstable results with respect to the relationship between standards-based teacher evaluations and student learning gains" (p. 17), even at the elementary school level where they predicted fewer implementation complexities than at other grade spans (e.g., multiple teachers at the high school level). They found challenges in the use of student learning data in

[4] Their review included evidence prior to the *Race to the Top* program or *NCLB Flexibility Waivers*, as well as evidence following those initiatives.

evaluation, such as the influence of the organization, grade span, assignment of students to teachers, subject area, and whether the student has multiple teachers. They noted how intensifying evaluation can impede school improvement efforts, and concluded that "the policy logic supporting this reform remains considerably stronger than the empirical evidence…alternative improvement strategies may yield more positive results and at a lower cost in terms of staff time and district funds" (p. 5).

In the review of the literature on this era's teacher evaluation and supervision systems few teacher perspectives emerged; Ritter and Barnett (2016) conducted one such study including interviews of almost 50 teachers, administrators, and policy makers. Seeking to learn lessons from schools and states implementing new evaluation systems, they found key components included meaningful feedback, frequent observations, clear explanations of rubrics, an environment of trust, ties to professional development, and feedback from peers (master teachers subject to evaluation by the same rubric). They concluded that "serious implementation of professional, rigorous, and comprehensive teacher evaluation systems represents a promising school improvement strategy" (p. 52).

Maine's Performance Evaluation and Professional Growth (PE & PG) Systems regulation (Chapter 180, 2015) directly requires or allows for each of the items Ritter and Barnett (2016) identified as keys to success. The Maine Education Policy Research Institute (MEPRI) conducted several studies and reported to the legislature annually during the development, piloting, and early implementation of Maine's PE & PG systems. First, Mason and Porter (2014) conducted a pair of surveys of Maine superintendents yielding approximately 75 responses in 2013 and in 2014. The survey results indicated strong belief that teacher effectiveness could be accurately evaluated, and indicated mixed support for linking student growth data to teacher

evaluation. More than three quarters of the participants were "very concerned" or "extremely concerned" about the analysis and interpretation of student growth data (p. 13). With a variety of other concerns emerging from the surveys, there was widespread support for an extended implementation timeline.

Between Mason and Porter's 2014 MEPRI report and Mason and Tu's 2015 MEPRI report, the legislature indeed acted to extend the timeline for piloting and implementing PE & PG systems. The 2015 report focused on the use of student growth data and teacher observations in PE & PG systems, drawing from case studies at seven school districts (interviews at each with superintendents or designees). These districts were purposefully sampled because they were farther along than others in using standardized student assessment data in PE & PG systems, and because they represented a variety of demographic characteristics. Mason and Tu reported that participants shared a variety of challenges in the use of student growth data (e.g., subjects less frequently tested, timeliness of assessment, validity) and a variety of ways in which those challenges had been addressed (e.g., leveraging partnerships and proficiency-based education efforts). Persistent challenges were noted with: a) identifying the teacher of record for a student's data without yielding perverse consequences such as a teacher narrowing their focus at the expense of school-wide support, and b) the challenge of using student growth data from the state-required test (the Smarter Balanced Assessment Consortium test, which ultimately was only given once in Maine).

Mason and Tu found in the same 2015 study that the seven districts sampled were each using a web-based teacher observation platform (e.g., iObservation, RANDA), and that each was overall satisfied with their observational system. Finally, the participants shared perspectives about progress toward useful peer feedback and connecting the PE & PG system to ongoing

feedback, support, and recognition. Mette and Fairman (2016) conducted the next MEPRI study, interviewing 11 administrators at four school districts that had received state approval for their PE & PG system plans. Those in the study shared successes, challenges, and lessons learned in piloting PE & PG components. Themes of success included a direction toward supporting individual professional growth, the use of online platforms to organize evidence and connect to resources, more clear professional practice standards, and the strengths of piloting before full rollout of system components. The uncertainty of shifting statute and regulation (especially with regard to student growth data) emerged in this 2016 report as it had in Mason and Tu's 2015 report. Other challenges that emerged were the time demands on administrators and the continued need for training and calibration for evaluators. Supervision gaps also led to requests for increased state funding; participants "see a need for increased instructional coaching in schools to support teachers and improve student learning outcomes" (Mette & Fairman, 2016, p. 18).

Finally, Fairman and Mette conducted the most recent (2017) MEPRI study, this time with 16 administrators in six purposefully-sampled Maine school districts with a variety of demographic characteristics. Relative to previous studies, they found that participants had made progress in understanding the professional practice models, realizing more consistency and depth, and more support for professional growth. Challenges remaining included the demands on time, questions on student growth data, and the continued need for calibration of evaluators. This 2017 study also reported on statewide data, analyzing Department of Education survey responses from 146 districts. These data showed such things as how districts chose to weight student growth data in summative evaluation (most chose between 20 and 39%), the types of evidence included in teacher evaluation (observations and portfolios were most common), little

use of merit pay, and how districts had used the $4,600 state grant for PE & PG systems (typically used for trainers, professional development, or devices to film instruction for peer feedback). Overall, participants sought to focus on professional growth rather than evaluation and appreciated the increased clarity of the new professional practice models.

Each of the MEPRI studies includes details beyond those summarized above, and each culminated in recommendations to the legislature and state. It is important to note that to this point the MEPRI research has focused on the perspectives of district and school administrators, which is similar to most of the literature across the United States in the era since *Race to the Top* and the *NCLB Flexibility Waivers*. Next, I describe the framework for my study which will also include the perspectives of teachers and supervisors.

Theoretical and Conceptual Framework

Current state regulation and several sets of theory help to frame the study. First, literature from educational leadership practice and theory defines what "teacher evaluation" means. Second, theoretical literature defines what "teacher supervision" means, specifically as this relates to professional growth within PE & PG models. Third, Maine Chapter 180 (2015) defines what is required of school districts with regard to Performance Evaluation and Professional Growth (PE & PG) systems, with the focus described above on summative evaluation. Hallinger, Heck, and Murphy (2014) describe such a policy theory underlying school improvement efforts focused on teacher evaluation. Fourth, with a theoretical lens that emphasizes teacher growth and teacher supervision, the well-regarded *SuperVision and Successful Schools* (Glickman et al., 2013) represents the growth-centered school improvement literature in contrast to Chapter 180's focus on evaluation.

Together, these elements serve as the conceptual framework I used to understand and analyze the PE & PG processes being implemented by Maine school districts between 2012 and 2017. I depict the elements below in a diagram showing: (a) through the overlap that evaluation-focused and supervision-focused approaches share some features, (b) that local districts face a tension between the evaluation-focused mandate and the supervision-focused improvement literature, (c) that school district implementation likely includes major evaluation and supervision efforts but likely does not fully encompass either approach, and (d) that school district implementation is inclusive of local assets and values and limited by local constraints.

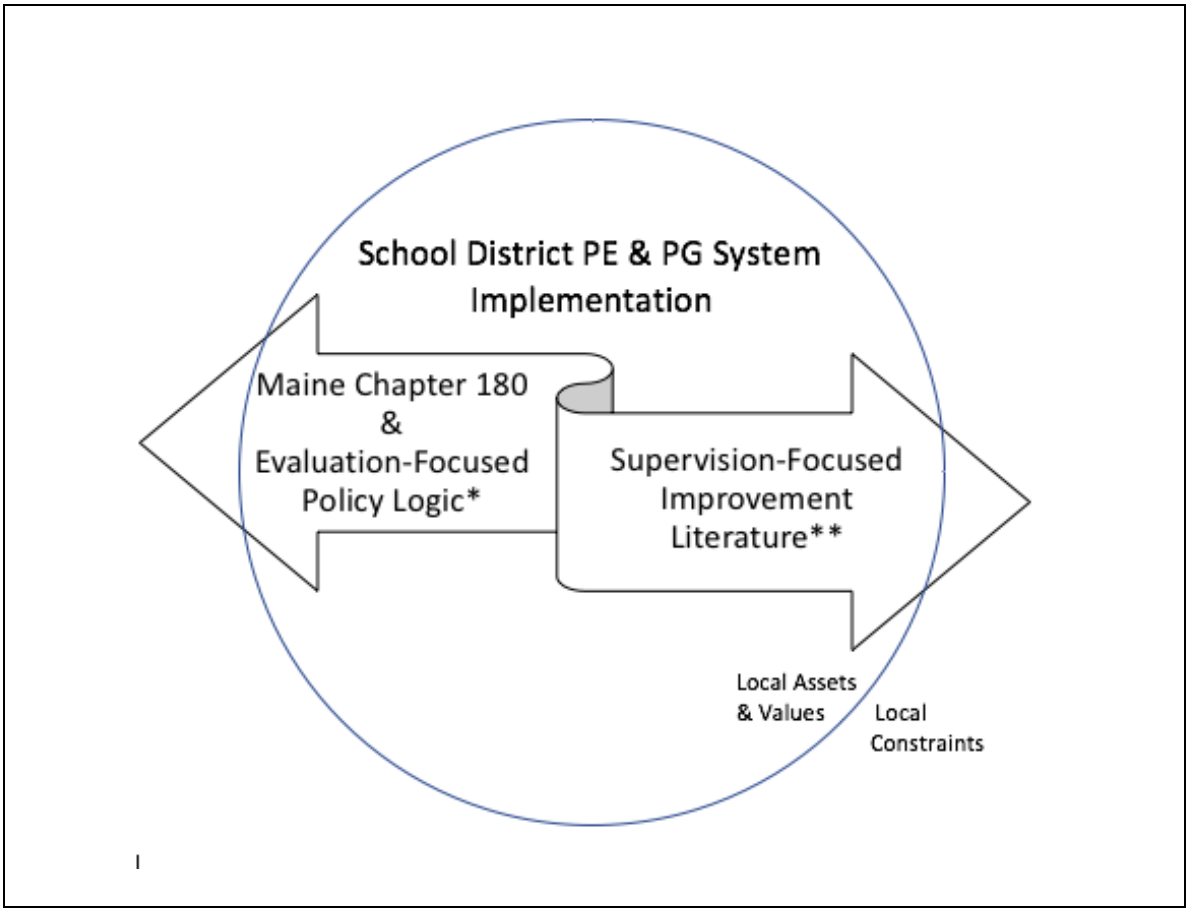

Figure 2.5. Conceptual framework for this study.
*e.g., as described by Hallinger, Heck, & Murphy, 2014
**As represented by SuperVision for Successful Schools (Glickman, Gordon, & Ross-Gordon, 2013)

I used this conceptual framework in analysis of the data to answer the research questions, describing the changes school districts have made in teacher performance evaluation and teacher professional growth, and factors that contribute to or arise as barriers to improved teacher effectiveness. Ultimately, in describing these changes and influencing factors, I describe how school districts adjusted in relation to the evaluation-supervision tension.

Prior to 2012, the state of Maine exercised very little influence in teacher evaluation and supervision. Now, Maine attempts to exercise significant influence on those functions yet also leaves much discretion in design and implementation to local districts. Maine's influence is potentially impacted by the tensions noted above and by an unpredictable future due to the *Every Student Succeeds Act* reducing the previous federal pressure for these mandates. In this study, I used Glickman, Gordon, and Ross-Gordon's "SuperVision" and Maine's regulatory requirements to understand and analyze local practitioners' perspectives of early local implementation of these Performance Evaluation and Professional Growth systems. I next describe the methodology by which I conducted this study.

CHAPTER 3

METHODOLOGY

As detailed in Chapter 1, the purpose of this study was to examine perspectives from the field regarding early local implementation of Maine's mandated multi-function systems of teacher performance evaluation and teacher professional growth (PE & PG systems). Specifically, the goals were to describe the perceptions and experiences of teachers, supervisors of teachers, and evaluators of teachers regarding major local changes in teacher performance evaluation and teacher professional growth, perceptions regarding whether and how local PE & PG systems were beginning to improve teacher effectiveness, and perceptions regarding factors contributing to or providing barriers to this improvement. To fulfill this purpose, I used a mixed-methods multi-site case study approach, with qualitative and quantitative data collection at eight school districts in Maine. It is important to note that the study took place during the 2016–2017 school year, when most school districts were either piloting or in a very early stage of implementing these complex systems. Thus, the systems may be different in later stages of implementation and more research is needed in coming years to accurately gauge how well the policy is meeting its intended goals.

In this chapter I lay out the methodology used in this study. I begin with research questions and definitions of key terms. Next I describe the study design, including instrumentation, recruitment of participants, data collection, management of data, and analysis of data. I conclude the chapter by describing biases I potentially bring to the study and my efforts to mitigate those biases to yield a trustworthy study.

Research Questions and Definitions of Key Terms

The study centered on three primary research questions examining the early local (school district) implementation of Maine's 2012 Educator Effectiveness law, 2014 Performance Evaluation and Professional Growth Systems regulation, and 2015 revisions to law and rule:

1. What do practitioners (teachers, supervisors, and evaluators) perceive as major changes in teacher performance evaluation and teacher professional growth in their school or district?
2. What factors do practitioners perceive as contributing to improved teacher effectiveness via teacher performance evaluation and/or teacher professional growth in their school or district?
3. What factors do practitioners perceive as barriers to improved teacher effectiveness via teacher performance evaluation and/or teacher professional growth in their school or district?

For the purposes of this study the terms used in the research questions above are defined as follows:

Practitioners at the local level means teachers, supervisors of teachers (e.g., Instructional Coach), and evaluators of teachers (e.g., Principal). *Teachers* means those who describe their primary role in school districts as providing instruction to students as a professional teacher in Maine. The term includes, for example, such roles as classroom teacher, physical education teacher, and reading interventionist. *Evaluators* means those who describe themselves as one of the people in school districts responsible for directly evaluating teacher performance for employment purposes. The term includes, for example, such roles as principal and director. *Supervisors* means those who describe themselves as one of the people in school districts

responsible for directly supporting teacher performance and growth, rather than evaluating that performance for employment purposes or providing direct instruction to students. The term includes, for example, such roles as literacy coach and instructional supervisor. Some school staff in Maine play multiple roles, which cause overlap between the functions of teaching, supervising, and evaluating. In this study, participants will self-report their primary function. *School Districts* or *Local* means the school district level (e.g., Regional School Unit, Municipal School District).

Teacher performance evaluation is a formal summative function focused on the organizational need for accountability, determining and documenting the level of a teacher's performance over a specific time period (Hazi & Rucinski, 2009; Platt et al., 2000).

Teacher professional growth is a formal or informal formative function involving both self-directed and guided teacher improvement; the process of guiding teacher professional growth is referred to as *supervision* (Glickman, Gordon, & Ross-Gordon, 2013).

Teacher effectiveness and *improved teacher effectiveness* in this study are as perceived and defined by the individual practitioners for the following reasons: (1) the term is not defined in Maine's PE & PG Systems regulation; (2) the local definitions, if they exist, would likely be unique to the professional practice model, locally-determined weighting of system components such as student growth data, and local values; and (3) even within the same school district educators' perspectives on effectiveness vary widely, drawing upon different logics and "intellectual, professional, and cultural histories" (Rigby, 2015, p. 378).

It is important to note that Maine's PE & PG mandates include teachers and principals; these research questions were investigated for teachers only. Also, these research questions were investigated in the piloting or early implementation phases of local districts' work to implement

the Performance Evaluation and Professional Growth Systems (PE & PG) mandates. Therefore the study is designed to capture changes that have happened and changes that are underway, along with participants' perspectives as to the potential for impact.

Study Design

This study uses a multi-site mixed method case study approach to yield rich data and in-depth descriptions about the phenomenon. The nature of a case study design is that it is not generalizable, however the site selection described below is such that the findings provide broad perspective about what is happening in Maine. Mixed-methods approaches are favored in educational policy studies to yield quantitative data on the present conditions and qualitative data on the consequences of policies (McMillan & Schumacher, 2010). This study of early local implementations (i.e., consequences) of a policy then appropriately uses a mixed-methods approach. The case study approach, "in-depth exploration of a bounded system based on extensive data collection" (Creswell, 2008, p. 476), is often used in policy research to generate understanding of complex situations (McMillan & Schumacher, 2010). In this mixed-method multi-site case study the case is defined and bounded by time (early implementation of statutes and regulations that changed in 2012, 2014, 2015, and 2016) and place (eight Maine local school districts).

In my approach with this study qualitative data obtained through semi-structured interviews (see instrument in Appendix A) aids in rich description and analysis of perspectives on experienced changes in performance evaluation and professional growth, the ways in which local PE & PG systems are or are not beginning to improve teacher effectiveness, and factors that have contributed to or raised barriers to improved teacher effectiveness. Quantitative survey data (see instrument in Appendix B) in this study aid in triangulation of those qualitative data via

descriptive and basic inferential statistics to describe how a more numerous pool of local practitioners experiences those changes and influences. Demographic data collected via both methods aids in identifying and describing patterns in respondent experiences or perceptions.

In the study design the qualitative and quantitative components were not dependent on the order of data collection. I utilized the following sequence (with some overlap): pilot and revise the quantitative instrument with a school district separate from the study sites, recruit participant sites, distribute the quantitative instrument via email, conduct interviews, email reminders to generate more survey responses, and conduct within-case and cross-case data analyses. Throughout I used analytic memos to record events, decisions, and thoughts that informed or influenced my findings.

Instrumentation

In addition to demographic information, the qualitative interview instrument (Appendix A) included a series of questions and probes to address each research question. I used this instrument to ensure during the interviews that each participant responded to each question, directly or indirectly. These questions begin broadly with major changes at the district, then narrowing down to specific changes in performance evaluation and in professional growth. Follow-up questions elicit perspectives about the perceived impact of the changes, local successes, local challenges, and the potential for improved teacher effectiveness. The final interview question asks participants whether professional growth or performance evaluation has received more local focus. The specific alignment of interview questions to research questions is as follows:

Table 3.1. *Mixed-Method Approach to the Research Questions*

Research Question	Qualitative Semi-Structured Interviews (Appendix A)	Quantitative Survey (Appendix B)
Research Question 1	Q1, Q4	Q21, Q31, Q32-37
Research Question 2	Q2a, Q2c, Q3a, Q3c	Q6-20, Q22-30, Q38
Research Question 3	Q2b, Q2c, Q3b, Q3c	Q6-20, Q22-30, Q39

In addition to demographic information, the quantitative survey (Appendix B) included item types in four total constructs aligned to the research questions as shown above: (1) changes in teacher performance evaluation, (2) changes in teacher professional growth, (3) factors positively or negatively influencing improved teacher effectiveness, and (4) changes in perceived teacher effectiveness. The first and second constructs utilized rank-ordering such that participants selected the five most major changes in their school or district. The options were drawn from a synthesis of the literature on teacher evaluation, teacher supervision, multi-function systems of evaluation and supervision, and Maine's regulatory requirements. An open-ended "Other" option was included so participants could describe changes different than those listed.

The third construct solicited participants' perceptions of performance evaluation and professional growth factors positively or negatively improving teacher effectiveness. These prompts were also drawn from a synthesis of the literature and utilized a Likert scale (Strongly Disagree, Disagree, Agree, Strongly Agree). Open-ended items were included for this construct as well to yield more data regarding influences on improving teacher effectiveness. Finally, the fourth construct in the survey solicited perceptions regarding changes in teacher effectiveness, again utilizing the Likert scale.

Piloting and Validation

To enhance the quality of the study, I piloted the quantitative survey with a sample district not to be used in the study. The pilot served two purposes: to ensure the survey could be completed in a reasonable amount of time (10–15 minutes), and to use validation strategies to improve the instrument. Through this pilot I identified that the survey was slightly longer in duration than intended, identified several ways to gain efficiency on checklist items regarding changes in performance evaluation or professional growth, and reviewed open-ended responses to be sure those constructs were written to elicit the type of data intended.

On scaled response items (i.e., Strongly Disagree / Disagree / Agree / Strongly Agree), I used Cronbach's alpha to determine consistency of answers on each construct and to aid in identifying any questions that should be eliminated to increase validity (McMillan & Schumacher, 2010). These results are shown in Tables 3.2 (professional growth construct) and

Table 3.2. *Professional Growth Alpha Results from Pilot Survey*

Item	Cronbach alpha
1. Teacher professional growth in my school district is focused on a shared vision of teaching and learning.	.798
2. In my school district, collaborative teacher professional growth is well-supported	.799
3. In my school district, individual teacher professional growth is well-supported.	.802
4. In my school district, teacher professional growth is well-supported within the school schedule / calendar.	.815
5. In my school district, teacher professional growth is well-supported outside of the school schedule / calendar.	.822
6. In my school district, non-evaluative staff are available to help teachers grow (e.g., literacy coach, instructional coach).	.824
7. In my school district, ongoing constructive feedback is provided to teachers.	.800
8. In my school district, teachers are supported in reflective inquiry.	.803
9. In my school district, evaluators (e.g., principal, director) are supportive of teacher professional growth.	.798

Notes. Overall Cronbach's alpha for the construct (n = 9) was .824. Each item was a Likert scale response (1 = strongly disagree, 2 = disagree, 3 = agree, 4 = strongly agree).

Table 3.3. *Performance Evaluation Alpha Results from Pilot Survey*

Item	Cronbach alpha
1. Teacher performance evaluation in my school district is focused on a shared vision of teaching and learning.	.817
2. In my school district, the amount of teacher time needed for the teacher performance evaluation process is reasonable.	.831
3. In my school district, the amount of evaluator time needed for the teacher performance evaluation process is reasonable.	.836
4. In my school district, classroom observations are frequent enough for accurate teacher performance evaluation.	.834
5. In my school district, clear standards exist for teacher performance evaluation.	.823
6. In my school district, clear rubrics or scales describe the standards for teacher performance evaluation.	.820
7. In my school district, the standards for teacher performance evaluation are relevant to teachers' responsibilities.	.821
8. In my school district, sufficient training is available for teachers to understand the process of performance evaluation.	.820
9. In my school district, sufficient training is available for evaluators to understand the process of teacher performance evaluation.	.830
10. In my school district, the process of combining measures into a summative effectiveness rating is clear.	.822
11. In my school district, the summative effectiveness rating accurately reflects teacher effectiveness.	.828
12. In my school district, the summative effectiveness rating distinguishes between levels of teaching effectiveness.	.831
13. In my school district, teachers were involved in developing the standards for teacher performance evaluation.	.829
14. In my school district, teachers were involved in developing the process for teacher performance evaluation.	.829
15. In my school district, teachers are involved in the ongoing governance of teacher performance evaluation (for example, "Steering Committee").	.825

Notes. Overall Cronbach's alpha for the construct (n = 15) was .836. Each item was a Likert scale response (strongly disagree, disagree, agree, strongly agree).

and 3.3 (performance evaluation construct) above. Based on the alpha scores for individual items and the overall total I retained all items in these constructs.

On open-response items I reviewed response types and conducted preliminary coding of the sample data with a second person experienced in coding open response items, to adjust questions and practices based on that collaborative review. As a result of using these multiple validation strategies, data gained through use of the quantitative survey instrument is viewed as valid for the purposes of this study. Below, I describe the sample and recruitment of participants. I then discuss the collection and management of data and describe how I analyzed the data to answer the research questions.

Sample and Recruitment of Participants

The sample for this multi-site case study is eight school districts in Maine (the district serving as the unit of analysis). I purposefully sampled the school districts to meet the following criteria: (1) a state-approved PE & PG system plan, (2) outside of my primary professional network in the Penquis region (to decrease the possibility that researcher familiarity may influence the data collected), (3) willing to participate as determined by the Superintendent of Schools or designee, and (4) a unique combination in this study of professional practice model and rural or non-rural status as defined by the National Center for Education Statistics (2016).

According to the Maine Department of Education's "Intent to Pilot" survey, 114 of 122 of responding school districts chose either the Danielson, Marshall, Marzano, or Maine Schools for Excellence (NBPTS/MSFE)[5] model for teacher supervision and evaluation (Lomonte, 2015); I sampled two districts that use each model. The National Center for Education Statistics (2016)

[5] The full description of this model is the "National Board for Professional Teaching Standards (NBPTS) Five Core Propositions and Indicators, along with the rubrics created by the Maine Schools for Excellence" (Chapter 180, Section 5, 2015).

classifies school districts in categories based on population and distance from urbanized areas; I sampled four rural and four non-rural districts. Thus, there are eight unique combinations:

Table 3.4. *Sites for the Multi-Site Case Study*

Professional Practice Model	Rural	Non-Rural
Danielson	Site 1	Site 2
Marshall	Site 3	Site 4
Marzano	Site 5	Site 6
NBPTS/MSFE	Site 7	Site 8

Site recruitment and site characteristics. I identified potential sites using the above criteria, PE & PG submission data from the Maine Department of Education, and an extension of my professional network and professional associations (i.e., colleague referrals, Maine Curriculum Leaders Association, Maine Education Association). After identifying potential sites I contacted key personnel to explain the study and obtain the permission of a gatekeeper (i.e., the Superintendent of Schools or designee) for the district's participation. While the nature of a case

Table 3.5. *Rural and Non-Rural Site Demographics*

Demographic Characteristic	Rural Sites	Non-Rural Sites
Student Population (Range)	300 to 1400	2000 to 3500
Student Population (Mean)	839	2473
Student:Teacher Ratio (Range)	7:1 to 21:1	11:1 to 13:1
Student:Teacher Ratio (Mean)	12.3:1	11.9:1
Free / Reduced Lunch Eligibility (Range)	22% to 67%	9% to 57%
Free / Reduced Lunch Eligibility (Mean)	47%	41%
Per-Pupil Spending (Range)	$8,100 to $17,500	$9,100 to $14,800
Per-Pupil Spending (Mean)	$12,988	$10,908

Note. Approximate range values are presented to maintain confidentiality.

study design is that the results are not generalizable, my variation-focused data collection approach provides broad perspective for Maine purposes and may be informative for states with similar characteristics. In Table 3.5 above I describe characteristics of the sample, using approximate values, ranges, and means to maintain the confidentiality of participants.

The rural sites in this study included those coded by the National Center for Education Statistics (NCES) as: Rural (Fringe), Rural (Distant), and Rural (Remote). Student population at the rural sites summarized in Table 3.5 above ranged from roughly 300 to roughly 1400 students (mean of 839), with approximate student-to-teacher ratios ranging from 7:1 to 21:1 (mean of 12.3:1). The rates of students eligible for free or reduced-price lunch at the rural sites ranged from approximately 22% to 67% (mean of 47%). Rural spending ranged from a low of approximately $8,100 to a high of approximately $17,500 per pupil (mean of $12,988).

The non-rural sites in this study included those coded by NCES as: City (Small), Suburb (Midsize), Town (Distant), and Town (Remote). Student population at the non-rural sites summarized in Table 3.5 above ranged from roughly 2,000 to more than 3,500 (mean of 2473). Approximate student-teacher ratios at the rural sites ranged from 11:1 to 13:1 (mean of 11.9:1). The rates of students eligible for free or reduced-price lunch at the non-rural sites ranged from approximately 9% to 57% (mean of 41%). Non-rural spending ranged from a low of approximately $9,100 per pupil to a high of approximately $14,800 per pupil (mean of $10,908).

Incidental to the selection process, several features further describe the overall sample: (1) seven of the eight sites served grades K-12 (the other site served grades K-8); (2) participants at six of the eight sites described their status as a pilot stage of PE & PG implementation (the other two sites described their status as implementing their systems); (3) the sites each served

Table 3.6. *Site Characteristics*

Site	Professional Practice Model	Rural Status, Grade Span	Stage of Implementation
Site 1	Danielson Model	Rural, K-8	Second Pilot Year (piloted previous year with Steering Committee)
Site 2	Danielson Model	Non-Rural, K-12	Implementing (after two pilot years)
Site 3	Marshall Model	Rural, K-12	Piloting
Site 4	Marshall Model	Non-Rural, K-12	Second Pilot Year
Site 5	Marzano Model	Rural, K-12	Second Pilot Year
Site 6	Marzano Model	Non-Rural, K-12	Second Pilot Year
Site 7	NBPTS Model	Rural, K-12	First Pilot Year
Site 8	NBPTS Model	Non-Rural, K-12	Implementing

between two and seven schools (mean of 4.75 schools per site, median of 5 schools per site); (4) the sites spanned central, western, southern, and coastal Maine (the gatekeeper at each potential northern Maine site ultimately did not choose to participate); and (5) one site was using a Teacher Incentive Fund grant (the TIF grants awarded money to districts and required certain evaluation system changes). I present the site characteristics in Table 3.6 above, omitting the number of schools and geographic location to protect participant confidentiality.

Survey and interview recruitment and sample characteristics. I worked with gatekeepers in the district to use or generate an email list of the district's teachers, supervisors of teachers, and evaluators of teachers, and to identify potential interview participants. I then emailed the quantitative survey to the email list, and recruited two to three participants from the site for the qualitative semi-structured interviews (some school districts do not employ a non-

evaluator supervisor of teachers). These interview participants met these criteria: (1) willing to participate, (2) knowledgeable about the district's Performance Evaluation and Professional Growth system (for example, a member of the district's PE & PG Steering Committee), and (3) filling one of the three roles at that site (teacher, supervisor of teachers, or evaluator of teachers). In some cases an interview participant was not found to meet a role (e.g., some school districts did not employ supervisors of teachers who are not also evaluators of teachers). In such a scenario I only interviewed two participants at that site. Thus, as shown in Table 3.7 below, I interviewed a total of 20 participants (2–3 participants at each of eight sites).

A few aberrations arose in the interview recruitment process that I proceeded with on the recommendations of the gatekeepers: (a) at Site 3 the gatekeeper assembled a group of evaluators to provide diverse perspective, (b) at Site 4 I was encouraged to interview two teachers who would bring different grade-span perspective, and (c) at Site 7 I was encouraged to interview two evaluators with different grade-span and historical perspectives. The total qualitative data include nine sessions with individual teachers, eight with individual evaluators of teachers, one with a group of evaluators, and two sessions with non-evaluative supervisors of teachers.

Quantitative surveys across the eight sites yielded a total of 366 surveys returned, approximately a 33% overall return rate (site-by-site return rates are shown in Table 3.7 below). During data analysis 64 surveys were found to have very few questions answered. Those survey responses were set aside leaving a total of 302 surveys in the analysis set. With a total population of more than 1100 at these sites, this was approximately a 27% return rate.

Table 3.7. *Site-By-Site Characteristics of Participants*

Site	Model (Rural Status)	Qualitative Interviews	Survey Response Rate
Site 1	Danielson (Rural)	Teacher Evaluator Supervisor	25%
Site 2	Danielson (Non-Rural)	Teacher Evaluator	27%
Site 3	Marshall (Rural)	Teacher Evaluators (group interview)	66%
Site 4	Marshall (Non-Rural)	Teachers (2) Evaluator	31%
Site 5	Marzano (Rural)	Teacher Evaluator	30%
Site 6	Marzano (Non-Rural)	Teacher Evaluator Supervisor	15%
Site 7	NBPTS (Rural)	Teacher Evaluators (2)	33%
Site 8	NBPTS (Non-Rural)	Teacher Evaluator	66%
Total		Teachers (9) Evaluators (9 interviews, including one with a group) Supervisors (2)	33%

Survey respondents in the sample identified themselves as follows:

Table 3.8. *Quantitative Survey Respondents by Grade Span and Role*

	Teachers n = 267 (88% of sample)	Evaluators n = 22 (7% of sample)	Supervisors n = 13 (4% of sample)
Elementary School Grade Span n = 122 (40% of sample)	107 (35% of sample)	7 (2% of sample)	8 (3% of sample)
Middle School Grade Span n = 83 (27% of sample)	73 (24% of sample)	7 (2% of sample)	3 (1% of sample)
High School Grade Span n = 97 (32% of sample)	87 (29% of sample)	8 (3% of sample)	2 (1% of sample)

Notes. Percentages may not total 100% due to rounding.

Interestingly, supervisors appear at more sites in the quantitative data than in the qualitative data. During site selection most site gatekeepers shared that they did not employ any non-evaluative supervisors of teachers, however at multiple such sites one or more individuals identified themselves with that label in the anonymous quantitative survey. I do not have any data to explain the difference between this self-identification and the gatekeepers' identifications.

Collection and Management of Data

As described, at each site I collected data from teachers, supervisors of teachers, and evaluators of teachers via qualitative semi-structured interviews and via quantitative surveys. In parallel, I collected some documents and artifacts relevant to the study: the state-approved PE & PG plan for each site, and (if provided by participants) other forms used in the PE & PG process. I reviewed the district PE & PG plans prior to interview sessions to prepare to understand participants' responses efficiently, and used these artifacts in data analysis when further information was needed to triangulate with the qualitative and quantitative data provided by

personnel. I additionally collected demographic information regarding the site and information about its implementation of the PE & PG system to use in description of each case and in cross-case analysis.

Qualitative interviews. The qualitative semi-structured interviews were typically 45-60 minutes in duration (see interview instrument in Appendix A). I provided a description of the study with risks and benefits of participation, and obtained informed consent along with permission to record the interview (see letter to participants in Appendix C). In parallel with my researcher notes, I audio-recorded the qualitative interviews, used an online transcription service, and verified the results to transcribe the data verbatim for use in qualitative data analysis software. Throughout, I used analytic memos to document the coding process, to document coding choices, and to begin understanding the case.

Quantitative survey. After validation and refinement of the survey instrument (described above; see Appendix B for the survey instrument) the survey took approximately 10-15 minutes for each participant to complete. While I took care not to collect any personally identifiable information, I electronically tagged survey results at each site to connect the results to the site. I distributed the survey to an email list of teachers, evaluators of teachers, and supervisors of teachers generated with gatekeepers at the site. I briefly described the study in the email contact along with risks and benefits of participation to inform consent; these were be provided in greater detail in the online survey.

Management of data. Following qualitative and quantitative data collection and transcription of interviews, I masked participant names (and other identifying information) in data analysis. The data and key were maintained in a password-protected environment; any paper documents were maintained in a locked environment. The key and any identifying

information will be destroyed within six months of completion and acceptance of this research by the University of Maine Graduate School. I loaded qualitative information into NVivo (qualitative coding software) for analysis; for quantitative data I iteratively used SPSS (quantitative analysis software), spreadsheet, and database software to aid in analysis.

Analysis of Data

Following data collection and initial data management, I systematically analyzed the data to answer the research questions. In Chapter 4, below, I present the findings within-case and cross-case.

Analysis of qualitative interview data. I used Eclectic Coding (Saldaña, 2016) in the first cycle of qualitative data analysis, simultaneously coding using methods purposefully selected to support the research questions and the nature of the data. Specifically, I used (1) Attribute Coding as a management technique for demographic and other basic characteristics; (2) Evaluation Coding to connect data to the research questions; and (3) Values Coding to unpack participants' values, attitudes, and beliefs related to the research questions. The provisional codes in Table 3.9 below were used as a starting point based on the scholarly literature. I used NVivo (qualitative coding software) to aid in the coding and analysis process. I expanded and revised these provisional codes based on the actual transcribed data. After first-cycle coding of several interview transcriptions, I reviewed the coding process then refined, finalized, and applied the codes to the transcriptions of each interview.

In the second cycle of analysis, I used Pattern Coding (Saldaña, 2016) to condense first-cycle codes into categories and sub-categories, later further condensing into themes and concepts. Using NVivo and a paper template I developed based on the research questions, I

Table 3.9. *Provisional Codes for First-Cycle Coding*

Attribute Coding	Evaluation Coding	Values Coding (examples)
Date of Interview	Change	V: (Values)
Personal	Change:PE	Accountability
Gender	Change:PG	Personal Responsibility
Professional	Teacher Effectiveness	Professionalism
Employment Role	Previous TE	Student Learning
School Student Pop.	Current TE	
Grade Span of School	Change:TE	A: (Attitudes)
District Student Pop.	Change:TE via PE	Consequences are overdue
Grade Span of District	Change:TE via PG	
Rural / Non-Rural	Contributing and Barrier Factors	Dislikes heavy regulation
PE & PG	+Factor: TE via PE	Hates mandates without resources
Stage of Implementation	-Factor: TE via PE	
Professional Practice Model (e.g., Danielson, Marzano)	+Factor: TE via PG	Wishes for merit pay
	-Factor: TE via PG	
	Recommendations	B: (Beliefs)
	Rec:PE	Process needs to be doable
	Rec:PG	
		Teachers need support

Note. These provisional codes for first-cycle Eclectic Coding (Saldaña, 2016) were generated based on the research questions and major themes from the scholarly literature.

gathered and analyzed identical and similar first-cycle codes (noting important words and phrases from the transcripts) to assign Pattern Codes. Additionally, I triangulated the second-cycle coding process with other relevant information from the site (e.g., documents). Subsequently, I analyzed each Pattern Code for use "as a stimulus to develop a statement that describes a major theme, a pattern of action, a network of interrelationships, or a theoretical construct from the data" (Saldaña, 2016). In final within-case and cross-case analysis, I triangulated analyses from the qualitative data with analyses from the quantitative data.

Analysis of quantitative survey data. For some items in the first two quantitative survey constructs, an option existed for the participant to generate a brief open-ended response (e.g., Other – Please Specify); in these cases I checked the responses for spelling prior to

analysis. The fourth construct contained open response item types. To enumerate data, I coded responses in this construct using keywords and the Evaluation Coding and Values Coding methods specified above; I additionally noted representative or outlier excerpts for use in communicating findings.

In order to address the research questions, I focused on descriptive statistics in quantitative analysis to describe patterns in responses. I used quantitative software (SPSS) to analyze these data, and also used spreadsheet software to generate descriptive statistics and identify representative or outlier responses for some question types.

In Chapter 4 I provide descriptive statistics within-case and cross-case for the individual and site demographics and for each survey construct. The descriptive statistics for these ordinal data include measures of central tendency and frequency tables. I scored individual questions as shown below in Table 3.10, with overall construct scores the mean of the individual question scores (for example, the professional growth construct score was the mean of nine individual question scores).

Table 3.10 *Scoring Applied to Individual Question Responses*

Scales:	Score:
Strongly Disagree (SD)	1
Disagree (D)	2
Agree (A)	3
Strongly Agree (SA)	4

In order to further inform analysis and understanding, I used basic inferential statistics (paired sample t-test, Kruskal-Wallis test with post hoc analysis) to determine if cross-case survey construct responses varied significantly based on role in school, professional practice model, or rural status. For all inferential statistics I used an alpha of .05 (Taylor, 1990). I did

not anticipate and did not find that additional inferential statistics (e.g., factorial ANOVA) were needed to answer the research questions.

Within-case and cross-case analysis. Ultimately, using the tools described above as well as paper templates and database software to organize the data, I triangulated the qualitative and quantitative results to generate comprehensive and rich description of what local practitioners experienced at each site and across the sites. I first conducted within-case analysis for each site, identifying site-by-site answers to the research questions through the qualitative and quantitative data. I next conducted cross-case analysis, seeking themes across many or all sites, by professional practice model, by professional role, and by rural status.

In the data analysis process attention was also given to response variation by grade span. The only prompt on which there appeared to be an aberration by grade span was whether a change occurred in professional growth via "increased instructional feedback through short classroom visits (a.k.a. 'mini-observations')." Very few elementary grades respondents across the sample identified this as a major change, while for middle grades respondents it was the third-highest reported professional growth change and for high school respondents it was a mid-level reported professional growth change. As the vast majority of indicators appeared not to show any pattern of variation by self-reported grade span, no further data analysis was conducted across those categories.

For the presentation of analysis and discussion I blended the qualitative and quantitative findings into a sequential organization by research question. I then compared and contrasted the findings with concepts from the theoretical and empirical literature as described in my theoretical framework. I include in the discussion limitations of the study and emphasis of significant findings and implications for researchers, policymakers, and practitioners.

Principal Investigator, Trustworthiness, and Researcher Bias

In this study I worked under the supervision of my doctoral committee led by Dr. Ian Mette (assistant professor of educational leadership). Dr. Mette has conducted multiple qualitative and quantitative studies in related areas, and has in-depth experience conducting research with human participants and has completed the human subjects training. Through the details I outline below and the guidance of Dr. Mette and my committee, I sought to maximize trustworthiness of this study and minimize bias.

In my current role as PK-12 Director of Curriculum, Instruction, and Assessment for a district of approximately 1,500 students, I have many day-to-day duties that include teacher supervision and teacher evaluation. In this role and preceding roles I have had numerous experiences with teacher supervision and evaluation over the course of my career that could have potentially impacted my collection and analysis of data. I briefly discuss those experiences here, along with the plans I implemented to bracket my biases. I also discuss why I perceive the current Maine context is a significant change.

Regarding teacher growth and supervision, I have personally benefitted from employment with a school district that has maintained support for teacher development even in the face of multiple large-scale budget challenges. This support manifested especially in collaborative professional goal-setting, financial support for conferences and coursework, financial incentive for increased educational attainment, and a sense of culture that values reflective practice and growth. Critical and supportive instructional feedback have been available to me from several colleagues throughout my career.

Regarding teacher evaluation, I have a history of frustration with evaluations not being performed by the building administrator in at least one local school, with a sense that this type of

choice is unfair to students and unfair to teachers. In my current role I am the lead administrator for the growth and evaluation process of a number of teachers, and collaborate with other administrators in the performance evaluation and professional growth process for several more teachers. I have experience with both full-class and mini-observations (Marshall, 2013). I played a primary role from February 2011 to present in crafting and implementing a new teacher PE & PG system for RSU #34, chair the Steering Committee for that system, and am responsible for conducting an annual audit of the instructional staff evaluation system for the district.

In formal preparation relevant to this analysis I successfully completed several courses in law and policy, varied coursework in teacher supervision and evaluation, varied coursework regarding school and district administration and leadership, multiple graduate-level courses in quantitative methods and analysis, and multiple graduate-level courses in qualitative methods and analysis. While I brought a variety of experiences and training to this proposed study, I also brought some biases that may impact the study. As introduced above, I believe support for growth and regular evaluations are critical to students and teachers. In the study this means I anticipated being more sensitive to open-ended responses that seem to limit support for growth opportunities or promote limiting or obstructing the evaluation function of PE & PG systems. Another potential bias I brought to this proposed study stemmed from my expectations for self and others. I view educators as salaried professionals and view professional growth as a responsibility shared by their school district, school, and the teacher. In this proposed study this meant I anticipated being particularly sensitive to open-ended responses that seemed to narrowly define the teacher's responsibility for ongoing professional growth.

A third area of bias, connected to the second, stems from a sense that teacher protections are very important (e.g., academic freedom, representation), but that these protections are

sometimes used to stifle any kind of change or to protect mediocrity. Thus, in this proposed study I anticipated being sensitive to open-ended responses that appeared to negate core teacher protections or offer protections well beyond protection from abuse. Finally, a fourth area of potential bias I brought to this proposed study stemmed from my view of the state's role in regulation. I brought to this study some frustration that what initially appeared to be a possibly helpful law (the 2012 Educator Effectiveness act) became loaded with many pages of regulatory detail, making it cumbersome to meet state requirements and (in my opinion) more likely that any employment actions based on evaluation would fail due to technicalities. As a whole, my perception of a large helpful change shifted to a perception of a large cumbersome state change. Thus, I made effort to be aware in the study that I would be sensitive to responses focused on bureaucratic details.

Addressing these known biases, and biases I did not anticipate, required awareness, analysis, careful design, and documentation. Through awareness, I prompted myself to look even more closely at the data relevant to my biases, and carefully triangulate with all available data. Careful design of instruments helped contain my biases, especially as aided by work with my chair to revise piloted instruments in the interest of clarity and fairness. Throughout the study, I noted thoughts and ideas that arose as I worked through data collection and analysis, and reviewed data and analysis with my chair and committee. I documented through excerpts the basis by which I made determinations and was sure to incorporate those excerpts – participants' voices – in the findings and discussion that follow. Finally, I reviewed the analysis and findings relative to the biases noted above to ensure that I was accurately treating the data.

CHAPTER 4

FINDINGS

The purpose of this study was to examine perspectives from the field as to major local changes in teacher performance evaluation (PE) and teacher professional growth (PG), perceptions regarding the ways in which local PE & PG systems are or are not beginning to improve teacher effectiveness, and perceptions of factors contributing to or providing barriers to this improvement. Data collection took place throughout the 2016–2017 school year, capturing perspectives as districts piloted or rolled out their PE & PG Systems under the new state mandate. Thus, the data may not reflect what systems would look like at a more mature stage of implementation. The sample included participants from eight school districts, each representing a unique combination of professional practice model and rural status; in total I conducted 20 interviews and received 302 survey responses.

For clarity of writing and to protect anonymity of sites and interview subjects, these terms will be used in describing the findings, regardless of the specific terms participants used:

- Maine's PE & PG laws (the set of laws and Department of Education regulations relevant to this topic)
- PE & PG System (e.g., evaluation system, TPEG, growth & evaluation system)
- Steering Committee (e.g., PE & PG committee, stakeholder group)
- Local Education Association (e.g., union, MEA)
- Administrative Team (e.g., leadership team, admin council)
- Peer Review (e.g., peer observation, colleague observation)
- SLOs / Student Learning Objectives (e.g., student growth data, growth data)

Where statistics are used here unless stated otherwise they are descriptive statistics, not inferential statistics.

In this chapter I present the findings that emerged from the qualitative and quantitative data collected in this study, arranged by research question. I begin by presenting the major changes underway in teacher performance evaluation and teacher professional growth (first within-case and then cross-case). I next describe factors practitioners perceive as contributing to improved teacher effectiveness, and factors practitioners perceive as barriers to improved teacher effectiveness. Finally, at this very early stage of policy implementation in Maine, I share participants' perspectives as to their PE & PG system's potential for improving teacher effectiveness. At the conclusion of this chapter I set the stage for discussion of the findings in Chapter 5.

PE & PG Changes Underway at Each Site

The first research question in this study is "What do practitioners (teachers, supervisors, and evaluators) perceive as major changes in teacher performance evaluation and teacher professional growth in their school or district?" I present major changes at each site in this section, with qualitative data in the forefront and noting any discrepancies found through triangulation with the quantitative data.

Site 1 is a rural K-8 school district using the Danielson model. At the time of the data collection Site 1 was in its second year piloting the PE & PG system, having piloted the previous year with only Steering Committee members. Major changes at Site 1 included a shift to the Danielson model, having previously used the certification standards for evaluation purposes. Site 1 is maintaining the three-year cycle it used previously and roughly maintaining the number of observations required each year. The observation structure remains similar to the previous

system: a modified clinical supervision model with the teacher providing written information, then an observation followed by a post-conference. SLOs are new to the site and are approached here in what interview participants describe as a Professional Learning Communities (PLC) style (Dufour & Eaker, 2009) focused on looking at strengths and lessons in the data. In this pilot year all staff are focusing on SLOs related to Danielson's Domain 2. Major changes also include efforts by the Steering Committee to calibrate with regards to the Danielson rubrics by using videos of one of their teachers, a volunteer for the role. Additionally, the Steering Committee is working to make concrete examples of what proficient and exemplary practice looks like. Concurrent with the rollout of the PE & PG system participants shared an increased focus on opening doors, professional conversation, and PLCs—overall putting professional goal-setting and growth in the forefront. The PE & PG system as a whole is more complex than the previous evaluation system, shifting from roughly a three-page document to extensive rubrics and either a digital or paper portfolio (at the teacher's discretion). Non-evaluative peer observation is also new to Site 1's PE & PG system and includes two elements: visiting a peer to observe something you wish to learn about, and having a peer visit to give feedback on something selected by you.

Site 2 is a non-rural K-12 school district using the Danielson model. At the time of the data collection Site 2 was beyond pilot stages, implementing the system by phasing teachers in as they concluded their three-year cycle under the previous system. The district is phasing in SLOs over the next three years. The district began PE & PG system changes prior to state law; it appears that the only components that solely stemmed from state legislation were the SLOs. Major changes in performance evaluation include a larger observation component, with peer observation in years 1 and 2 of the cycle and administrator observation at least once in year 3 (using the clinical supervision model). Site 2 is shifting to a different technology platform to

manage evaluation system documents, in hopes it will be less cumbersome. The district pooled professional development funds at the district level; created rubrics for review of proposals; and began an internal grant system where teachers write grants to fund professional development proposals including compensation for time. This was part of a shift toward an internal professional development focus and away from most external workshops or conferences. The district added three teacher professional development days, one of which is up to individual teachers' discretion (they provide some documentation, and collaboration is encouraged).

Site 3 is a rural K-12 school district using the Marshall model. At the time of the data collection participants described Site 3 as in a pilot stage of implementation (though many components had been in place for years). The district undertook a major change in 2010 before the state law took effect, shifting to a three-track model: Collegial, Professional, and Administrative. The administrative track contains the summative evaluation, and evaluators can keep teachers on the administrative track for multiple years if needed. The collegial and professional tracks, which interview subjects seemed to use interchangeably, allow teachers to think more about what they want to focus on and learn about. The teacher interviewed shared that the new system is more teacher-driven, allowing for more flexibility and interdisciplinary team work. The evaluators interviewed shared that the district also focused on increasing funds to provide opportunities for teachers in the collegial and professional tracks, such as attendance at conferences and supporting teachers in pursuing a Master's degree; in contrast the teacher interviewed perceived that professional development funds were refocused internally, with less support for outside workshops. New, more detailed rubrics replaced the preceding rubrics and teachers were trained on the new rubrics. Evaluators used a videotape of one of their teachers to calibrate evaluation. As part of the 2010-era change, observation practice shifted from the

clinical supervision model once a year (with teachers focused on how many "Commendables" they earned) to a walk-through model. The direct influences of the state PE & PG Systems law at Site 3 were the implementation of SLOs and averaging ratings into a summative evaluation.

Site 4 is a non-rural K-12 school district using the Marshall model. At the time of the data collection Site 4 was in the second pilot year of their PE & PG system. Those interviewed noted large changes from the preceding system. The preceding system was primarily observation-based, with a minimum of one observation (usually bell-to-bell) annually on a three-year summative cycle. The new system maintains a three-year cycle, and preceding efforts to open doors and build a peer review culture will continue. Teacher goals existed before the new PE & PG system, but were not directly part of the system. Now, goal-setting is much more tied into the system along with self-evaluation; goal approval; more frequent and often shorter observations; and the inclusion of artifacts, evidence, and student growth into a teacher's summative evaluation.

Site 5 is a rural K-12 school district using the Marzano model. At the time of the data collection Site 5 was in its second pilot year. The district shifted from an evaluation model based on Danielson to the Marzano model. Major changes to the PE & PG system include the addition of iObservation and having multiple administrators involved in each teacher's evaluation. The previous and current systems each used a three-year cycle, but in contrast to the previous system's single clinical supervision observation every three years there are now intended to be a total of nine brief observations (20-minute informal, three per year). Years 1 and 2 on the current system are non-summative (e.g., professional learning, peer feedback, peer observation, instructional rounds, collecting evidence). The third year of the cycle includes a reflection on goals and a summative evaluation. If the teacher and evaluator agree on evaluation ratings the

evaluation is turned in; if not there are several more months of evidence collection. Major changes to professional growth were underway prior to the change in state law: a focus on instructional leadership, peer observations, early release times, instructional rounds including teachers, PD efforts regarding active pedagogy, and project-based learning. Time has been allotted to administrator training and calibration sessions, to the existence of a Steering Committee, and teacher training for iObservation.

Site 6 is a non-rural K-12 district using the Marzano model, in its second pilot year at the time of data collection. Site 6 changed its professional practice model to the Marzano standards and introduced use of iObservation. This shift included significant professional development about the elements of the new system, in parallel with professional development on shifts toward customized learning, reflective practice, and associated technological tools. All three interviewed intertwined the PE & PG system shifts with these other shifts, sharing that the district had become more organized and purposeful, shifting to better articulation of "look-fors," which they found more meaningful and more objective. Calibration efforts with the Marzano standards seem to have focused on the administrative team seeking inter-rater reliability, generating examples of proficient or higher-level practice. With the shift to Marzano and parallel efforts, the lens expanded to look more at what students are doing in the classroom in addition to teacher efforts, and the culture shifted toward more transparency and welcoming of classroom visitors. Each interviewed shared that more attention is now paid to student evidence of learning, which is newly included in the evaluation system and has resulted in more data-driven learning team time and more teacher thought given to formative assessment. Other changes at Site 6 included a shift from a three-year summative evaluation cycle to evaluating all teachers in a single year, and an increase from typically one clinical supervision observation on

alternating years to multiple and less in-depth "walk-through" observations each year, with the option of additional formal observations. Professional growth efforts include sharing with teachers how to "lift" to the next level on iObservation, an expansion in district staff dedicated to instructional coaching, and Steering Committee work to explain the Marzano system components, design questions, and elements. The quantitative survey data differed in two aspects from the qualitative data: (1) Few respondents noted "Increased training about professional practice standards" as a major change in the district, and (2) numerous respondents noted "Increased teacher professional goal setting" as a major change in the district.

Site 7 is a rural K-12 school district using the NBPTS model. At the time of the data collection Site 7 was in its first pilot year, having developed the system the previous year; the development and rollout timelines were impacted by multiple administrative turnovers. The site changed quickly from the previous system to the state model system; interview participants indicated that the anticipation of state approval was a primary consideration in this choice. With the state model system, many elements were new to the site: the standards, the process (e.g., goal-setting, portfolio), and SLOs. Some administrative calibration efforts joined the rollout of the new system, along with teacher professional development on the standards, rubrics, and goal-setting elements of the new system. At the time of the interviews some elements had been rolled out to teachers, while others (e.g., peer observation) had been discussed but not yet initiated.

Site 8 is a non-rural K-12 school district using the NBPTS model; the site has been part of a federal Teacher Incentive Fund (TIF) grant. Site 8 was the only site in the study that used merit pay directly connected to the summative evaluation ratings. At the time of the data collection Site 8 was several years into implementation of the TIF model. Under TIF the district

"started from scratch", developing an entirely new PE & PG system with new standards, technical platforms for managing the system, SLOs, peer feedback, evaluator calibration, trainings, incentive pay, and methods of collection of evidence to show proficiency in instructional or "off-stage" (e.g., parent communication) standards. Performance evaluation elements that shifted also include an increase in observational frequency. Action plans for those required to show improvement under the system are now more detailed and more supported (e.g., readings, lesson plan templates, lesson plan feedback, multiple observers, multiple evaluators involved with observation, weekly support meetings). Major changes in professional growth include more detailed feedback, bringing in external trainers and consultants for elements of the PE & PG system, funding for external conferences, the opportunity for teachers to participate in compensated resource groups (e.g., group readings, study groups), trainings for teacher-facilitators of those groups, and other trainings specific to the TIF system.

Overall Changes in Teacher Performance Evaluation and Professional Growth

Above I shared site-by-site changes underway at the school districts in this study, in regards to the first research question: "What do practitioners (teachers, supervisors, and evaluators) perceive as major changes in teacher performance evaluation and teacher professional growth in their school or district?" In this section I synthesize those site changes, blending in patterns noted by professional practice model, rural status, and professional role. Analysis of major changes shared by interview and survey respondents identified themes primarily affecting performance evaluation, primarily affecting professional growth, and especially more broad themes of changes that affect both performance evaluation and professional growth. At all but one site the survey responses (see Tables 4.1 and 4.2) about

Table 4.1. *Performance Evaluation Changes (Count)*

	Changed professional practice standards.	Developed rubrics or scales for professional practice standards.	Increased training about professional practice standards.	Increased training about the evaluation process.	Increased frequency of full-class observations.	Increased frequency of "mini" or "walkthrough" observations.	Increased frequency of unannounced observations.	Increased frequency of pre-planned observations.	Increased frequency of summative evaluations.	Increased number of people with evaluation responsibilities.	Increased use of evaluation outcomes in collaboratively-set professional goals.	Increased use of evaluation outcomes in evaluator-set professional goals.	Increased use of favorable evaluation results for opportunities a person would typically desire.	Increased use of unfavorable evaluation results for consequences a person would typically not desire.	Increased use of student growth data in evaluation.
Danielson Rural	1	***5***	***3***	***3***	1	1	1	0	0	0	***3***	1	0	0	***5***
Danielson Non-Rural	6	***10***	7	***12***	2	***10***	1	3	4	7	5	2	0	0	***13***
All Danielson	7	***15***	10	***15***	3	11	2	3	4	7	8	3	0	0	***18***
Marshall Rural	***10***	***14***	1	3	1	***10***	2	4	1	1	6	3	0	0	***21***
Marshall Non-Rural	9	***33***	5	14	0	***18***	2	1	0	4	2	2	0	1	***20***
All Marshall	19	***47***	6	17	1	***28***	4	5	1	5	8	5	0	1	***41***
Marzano Rural	***15***	***12***	7	8	5	***13***	7	1	1	0	2	1	0	1	4
Marzano Non-Rural	10	***11***	3	8	2	***15***	3	2	4	2	5	8	1	5	***19***
All Marzano	***25***	***23***	10	16	7	***28***	10	3	5	2	7	9	1	6	***23***
NBPTS Rural	***7***	***10***	3	3	0	1	1	0	0	0	0	0	0	2	***5***
NBPTS Non-Rural	***29***	***36***	22	21	7	10	6	6	9	19	8	9	4	7	***44***
All NBPTS	***36***	***46***	25	24	7	11	7	6	9	19	8	9	4	9	***49***
All Respondents	***87***	***131***	51	72	18	78	23	17	19	33	31	26	5	16	***131***
All Evaluators	***10***	***10***	4	2	2	6	6	0	2	2	1	1	0	1	***10***
All Supervisors	3	***4***	1	***4***	2	***4***	0	0	1	2	2	1	0	0	***5***
All Teachers	***74***	***117***	46	66	14	68	17	17	16	29	28	24	5	15	***116***
All Rural	***33***	***41***	14	17	7	25	11	5	2	1	11	5	0	3	***35***
All Non-Rural	54	***90***	37	***55***	11	53	12	12	17	32	20	21	5	13	***96***

Notes. Survey participants each noted the top three changes they perceived. The most frequent responses within-group are bolded and italicized. Few participants chose "Other – specify;" those data are addressed along with other open response items.

Table 4.2. *Professional Growth Changes (Count)*

	Increased funding for workshops or conferences	Increased funding for university coursework	Increased teacher support via coaching	Increased teacher peer support and/or peer collaboration	Increased teacher professional goal setting	Increased support for teacher professional goals	Increased time for professional growth embedded into the school schedule (e.g., early releases, in-service days, within meeting schedule)	Increased instructional feedback through short classroom visits (a.k.a. "mini- observations")	Increased instructional feedback through full-class visits / observations	Increased use of a professional development software system	Increased training and/or collaboration on analysis of student data	Increased emphasis on reflective inquiry	Differentiated professional development
Danielson Rural	0	1	1	***5***	***5***	2	3	0	0	0	***4***	1	0
Danielson Non-Rural	1	2	5	***12***	***10***	4	9	3	0	***10***	6	3	***10***
All Danielson	1	3	6	***17***	***15***	6	***12***	3	0	10	10	4	10
Marshall Rural	4	3	1	7	***13***	***10***	***15***	5	1	0	7	4	6
Marshall Non-Rural	0	0	***11***	4	***20***	5	7	***9***	3	2	5	4	3
All Marshall	4	3	12	11	***33***	***15***	***22***	14	4	2	12	8	9
Marzano Rural	1	2	3	5	***9***	1	***12***	***9***	3	2	3	6	7
Marzano Non-Rural	0	1	***13***	5	***16***	5	8	8	2	***10***	8	5	6
All Marzano	1	3	16	10	***25***	6	***20***	***17***	5	12	11	11	13
NBPTS Rural	1	0	***3***	***3***	***5***	***3***	***4***	2	0	0	***3***	0	0
NBPTS Non-Rural	9	1	14	***28***	***36***	17	***33***	4	10	14	7	8	19
All NBPTS	10	1	17	***31***	***41***	20	***37***	6	10	14	10	8	19
All Respondents	16	10	51	***69***	***114***	47	***91***	40	19	38	43	31	51
All Evaluators	0	0	4	***9***	***9***	5	5	4	2	2	2	0	***6***
All Supervisors	0	1	3	***5***	***4***	1	***6***	1	1	1	2	1	1
All Teachers	16	9	44	***55***	***101***	41	***80***	35	16	35	39	30	44
All Rural	6	6	8	***20***	***32***	16	***34***	16	4	2	17	11	13
All Non-Rural	10	4	43	***49***	***82***	31	***57***	24	15	36	26	20	38

Notes. Survey participants each noted the top three changes they perceived. The most frequent responses within-group are bolded and italicized. Few participants chose "Other – specify;" those data are addressed along with other open response items.

major changes in the district paralleled the qualitative interview data. With the exception of tangible increases in support, which occurred only at non-rural sites, findings regarding changes underway did not vary by rural status. Similarly, only one difference was noted by professional role in responses regarding changes underway; this was a small distinction and is described below. Across all roles and sites three major clusters of themes emerged: (1) increased techno-rational ideology of evaluation, (2) shifts from the traditional observation and supervision paradigm, and (3) increased goal-setting and focus on professional growth.

Increased Techno-Rational Ideology

Across the sites in this study, major changes were underway to implement new structures approved by the state that defined effective teaching and provided a framework for evaluation and growth. As one evaluator shared, "The standards are new; the process is new." Specifically, these changes included professional practice standards, detailed rubrics, and processes for incorporating student growth data into evaluation systems. At three quarters of the sites survey respondents largely agreed that "teachers were involved in developing the standards for teacher performance evaluation" and that "teachers were involved in developing the process for teacher performance evaluation." While these elements or their structure were selected or adapted via a process of shared governance, the options available were limited by state law and regulation. Thus for the sites in the study policymakers at the federal level (through pressures and incentives) and at the state level (through legislation and regulation) had used a techno-rational policy approach (Webb & Gulson, 2013) to newly define effective teaching.

Professional Practice Standards. At each site in the study the school district adopted a wholly new set of professional practice standards (i.e., Danielson, Marshall, Marzano, or NBPTS). This change was raised in each site's qualitative interviews and was consistently

raised in survey data as well (survey data for this construct are presented in Tables 4.1 and 4.2 above). In a few of the districts the change in standards had begun prior to Maine's 2012 PE & PG Systems law, but the law prompted the shift in standards at the majority of the districts. There was no pattern noted in why districts chose the Marshall or Marzano models, but at both Danielson sites interview subjects shared that teachers primarily selected the model: "We researched what was going on nationwide and we came up with Marzano, the national teacher boards, and Danielson…Danielson was the one that K-12 teachers overwhelmingly selected."

At both NBPTS sites in the study external factors heavily influenced the choice of model. At one of the those sites TIF grant access was the motivation: "The reason why we did the TIF grant was not because we thought the whole system was a brilliant idea; it was more so we were in desperate need of professional development money." At the other NBPTS site an imminent deadline and anticipation of state approval influenced the decision: "When the legislation got passed, we did not make any steps to fall in line…Then we did a pretty rushed job last year because we had some superintendent challenges and turnover, and so they just never started the process."

Implementation of the New Standards. New rubrics or scales to describe and implement the professional practice standards were another major change at all eight of the sites, supported by interview and survey responses. Maine's PE & PG laws require rubrics for each standard; the creators of the professional practice models in this study each provide rubrics for the models' standards. These new devices were described as much more detailed than what was previously in place. "We have a series of new rubrics…that are much more in-depth than in the past…It's much more effective because it breaks down…what a three or a four looks like."

At half of the sites interview responses also indicated a change in technical systems related to the PE&PG system; some of these systems focus on evaluation components and some include evaluation and professional growth components. Both Marzano sites adopted the iObservation system; an evaluator described the embedded professional growth aspects of that system:

> Let's say you've got a developing on something and you're thinking, 'How do I lift to the next level? As I talked to [teacher], we discussed it. I'm going to go and take a look.' And you can just type in and you're going to have a short three minute video that comes up. They're great resources.

At almost all sites where technical systems changed, interview and survey open responses indicated allocation of professional development time to training staff on the technical system.

Beyond the technical systems, survey responses and interview data indicated an increase in training about professional practice standards and about the evaluation process at most of the sites. At a majority of the sites calibration to the professional practice standards was conducted for the administrative team (and sometimes the whole Steering Committee); this occurred at both Marzano sites, both NBPTS sites, at one of the Marshall and at one of the Danielson sites. An evaluator shared that their administrative team videotaped a volunteer teacher, then "we sat down as a group and had conversation using the rubrics to assess that teacher so that we're all on the same page." At another site this work was described as more extensive:

> Administrators have been doing a lot of unpacking of Marzano's framework and looking at what this would look like in a classroom at a four? At a three? Being transparent with teachers about that. These are the look-fors. Here's what we are hoping to find. The

steering committee that I'm a part of, we do a lot of gathering of resources to help explain the components, the different design questions and elements.

At six of the eight sites (both Marshall, both Marzano, both NBPTS sites) interview subjects actively described the new PE & PG system as more evidence-based than their previous systems, drawing in such things as portfolios, student work, and other "off-stage" artifacts that could not be seen in a classroom observation. At the Danielson sites, interview subjects spoke of such evidence sources but there was less emphasis on an increase in focus on evidence (perhaps because both Danielson sites had previously used a recognized set of standards in growth and evaluation work).

Student Growth Data. The "increased use of student growth data in evaluation" was another change noted across all eight sites. The approach varied across the sites; regardless of whether the approach exactly matched the state's Student Learning Objectives (SLO) model most participants used the "SLO" acronym to describe the use of student growth data. Implementation of SLOs was raised in qualitative interviews at all sites; at each either the new state laws or TIF grant had prompted use of SLOs.

Beliefs and values regarding Student Learning Objectives were frequently raised by interview and open response subjects. Unlike most factors, SLOs arose in both the constructs about factors contributing to increased teacher effectiveness and about factors as a barrier to increased teacher effectiveness. Some participants were optimistic about the potential for SLOs to benefit teacher effectiveness, such as an evaluator who shared:

> The way we structured our SLOs is such that hopefully by design teachers are creating up front, they're investing the time, energy and effort into creating an SLO that is meaningful, and that their focus can hopefully be on, what is my impact on their learning,

> and taking some deliberate steps to influence student achievement. The feedback I got from the people who participated in the pilot last year was really promising in that regard. … As we layer in those SLOs, if we can frame it that way for people and have it not be a high stakes thing, but a learning thing, I think there's a lot of opportunity there.

At some sites, SLOs were being piloted or implemented in a "Professional Learning Communities" style (Dufour & Eaker, 2009) focusing on collaborative work with assessments and data. Other approaches included the state model approach to SLOs, variations on that model, or in one case pushing the inclusion of SLOs out multiple years via a phased-in pilot in the hopes that state law or regulation would change and the district could avoid including that piece in teacher evaluation and growth systems.

Overall, the data from those in all roles trended solidly toward the negative with regard to Student Learning Objectives, with interview and survey participants speaking candidly about their views. One teacher referred to SLOs as "the scariest piece…to everyone," sharing challenges with teachers who have very small numbers of students (e.g., interventionists) or very large numbers of students (e.g., arts teachers). A teacher at another site had skepticism about validity, "I can pick a topic where my students will fail the pretest and successfully pass a posttest – big deal." At a third site, a teacher at a different grade span shared a similar thought: "The actual SLO piece is forced upon us, according to law, but is a game that anyone can win; it's a joke." An evaluator described her site's challenge with an example:

> Think of chemistry. If you've never taken chemistry before how are you going to score on a pretest? Then at the end, you're going to obviously know a lot more than you used to. You're going to show extreme growth. We've tried to come up with all sorts of ideas to get around that margin of growth really sort of impacting the SLO scores, but it's nearly

> impossible. My reason for saying the SLOs, I wish they would just go away, is just because I think that it's really hard to do them with accuracy and well, so they just cause way more problems then they're worth. The time it takes to do them kills me. When I think of all the other great things you could do with that time, it kills me.

A supervisor at a different site shared similar thoughts about the time investment in SLOs:

> I felt overwhelmed and a little angry with the SLO's as well because I did those with my students and I found that ... I don't mind, I'm always progress monitoring. I felt like I understand the purpose of it but I'm not really sure it's going to change some teachers… Sometimes it's not about improving instruction it's about getting all your things in place. As a literacy coach and seeing where we could really build some awesome units, we're not doing that. We're spending our time getting these other pieces done. I've been frustrated with that.

Finally, an evaluator, while generally positive about her district's approach, summarized her thoughts concisely at the end of the interview, "If I wasn't clear, I think SLOs are stupid." Overall, the data in this study show that districts' shifts to include SLOs in teacher PE & PG systems arise from policymakers' objectives rather than from local initiative.

Shifts in Supervision, Observation, and Culture

The second cluster of major changes that emerged in the data regarded shifts away from teacher isolation. This manifested in a variety of ways such as more discussion of instructional practice and more frequent visits to classrooms by evaluators, supervisors, and peers. This opening of doors had multiple intended purposes, including increase of evaluation accuracy, increase of constructive feedback, and contribution to professional dialogue about teaching

practice. As will be shown in a later section, these intended outcomes were not consistently realized.

Expansion of the observation model. At most of the eight sites the prior teacher evaluation system required a single observation of teacher practice either annually or once on a multi-year cycle. In most cases, this observation was a full-class observation following the clinical supervision model. "What we were using was the canned model where you make the appointment, you do the pre and post conference, and it's a one shot deal." At three-quarters of the sites observation frequency increased with the PE & PG system changes, and the observation model shifted at nearly all of those sites to include more brief observations (typically 10–20 minutes). The addition of brief observations occurred at both Marshall sites (Marshall favors ten brief observations with feedback per year), both Marzano sites (the Marzano model includes brief observations with data collection on iObservation) and was reported as a major change by survey respondents at one Danielson site.

This inclusion of brief observations, however, does not necessarily mean replacement of the clinical supervision model; many of the sites retained a clinical supervision model observation as an optional or periodically required component of the PE & PG system (e.g., on the summative year of a three-year cycle). Sites also did not report adding evaluators, though in some cases evaluation responsibilities were added to existing positions or used as a rationale to maintain an administrative position in budget cut discussions. Therefore this increase in observational frequency for evaluation purposes appears to come from the pre-existing administrative staff. A supervisor shared, that "it feels like more is being asked of the people in those positions. A lot more. [Principal] and [Asst. Principal] are less available because they … are busy with the amounts of observations and keeping track of our work." In the data there did

not appear to be a pattern across sites as to whether the additional observations were typically pre-planned or unannounced.

Peer review. Classroom observations for growth or feedback purposes need not necessarily be administrative, however. The Maine PE&PG Systems law requires a peer review component, which was reported as a major change at five of the eight sites (both Danielson sites, and one each of the sites using the other models). Interview subjects at each of the three other sites discussed peer components, but the sites had not yet launched that change. In all cases this peer component was done or planned with deliberate separation from summative evaluation. At some sites the described peer components were fairly informal, such as "visit a teacher, for something that they were interested in getting better at, and also invite a teacher into their classroom to help them improve in an area." At one site certain teachers were trained to give peer feedback and compensated for their work visiting classrooms and documenting the feedback.

Cultural efforts. In interview data a pattern emerged at half of the sites (including both Danielson sites) about deliberate overall efforts to build cultures with more "open doors" and more transparency. A supervisor described her part in that district-wide effort:

> My role is to see what staff are doing well so that I can share their successes with others because teachers have always worked in isolation. So they don't recognize what they're doing as successful because they haven't seen otherwise. So I go into classrooms to observe something and say, "Oh, you know who did that strategy really well?" And to try to direct people to utilize each other. It's really what's important, I think.

Several interview subjects shared that the cultural shift is gradual with some teachers still uncomfortable with classroom visitors. A teacher shared,

> It's that open-door. In and out, in and out, you have people asking you what you're doing, why you're doing it, which I think is a shift in general, in a positive way, because these are the what's, why's, and how's are the things that we're asking our students to think about, and now we're kind of doing it, and people are out of their comfort zone, which is a good and a bad thing, and it has given, at least for me, a lot of opportunity to learn from a lot of teachers that I would probably have never had direct contact with.

A teacher at another site shared deliberate efforts to yield similar outcomes:

> We're asking former or current peer evaluators to open their classrooms so that we can do some on-the-spot training, so that teams can come in and just sit and see what it's like. We would like as many people as possible to be involved with the observation because as I said, that's up to four touches, really examining what's good instruction, what's effective instruction.

Finally, an evaluator shared that while she encourages teachers to visit and learn from others, few teachers take advantage of the opportunity:

> I do encourage teachers to go and observe another classrooms. Even other buildings. Do they take advantage of that a lot? I would say no. And we just had this discussion last week at our staff meeting. And I brought up the fact again that this is been offered. I will get support in your classroom. I'll get coverage. You may go look. The hard part is that teachers don't want to leave their classroom. They want to teach their kids. So that's kind of that tug and pull. It sounds good, they want to get out and go look at someone that is doing maybe more of an applying or innovating job versus a developing…then they get into the routine.

Overall these efforts to shift culture were viewed positively by participants. In a later section findings will be described as to how such cultural efforts were perceived to contribute to increased teacher effectiveness.

Increased Focus on Goals and Growth—With Mixed Support

The third and final cluster of major changes at local school districts involved an increased focus on teacher goals and teacher growth. The setting of goals and focus on supporting growth was in the forefront for many participants, to a greater extent than evaluation and accountability. However, the resources for that support did not necessarily follow.

"Increased teacher professional goal setting" was one of the top survey responses about professional growth changes at every site in the study, and was raised as a major change in interviews at nearly half of the sites. A teacher was clear about the focus for her site: "Our goals are all directed toward student learning and outcomes. Better instructional practices means better student outcomes." However, "increased support for teacher professional goals" and related qualitative responses were much less frequent, a middle or low-frequency response at six of the eight sites. Responses by role largely paralleled each other, but across all sites teachers and supervisors were somewhat more likely than evaluators to share that time had increased for "professional growth embedded into the school schedule."

Across all reported information tangible increased support, beyond a change of focus or change of administrators, was found at three sites. The non-rural Danielson site shifted some professional development funds to compensation for collaborative groups and added multiple teacher release days to the calendar. The non-rural Marzano site increased supervision (e.g., coaching) at the district level in parallel with the PE&PG system shift. The non-rural NBPTS site also added compensation for collaborative teacher groups and the resource of a temporary

TIF coach. Only this site appears to have significantly increased professional development funding support for external conferences and workshops; these funds, the funds for the coach, and the compensated collaborative group funds were provided through the TIF grant.

Notably the three sites with tangible increases in support were each non-rural. While some rural sites were refocusing efforts toward increased support for teacher growth none shared such tangible increases in funded support. Relevant challenges of funding and geography are discussed in more detail in a later section.

Focusing on professional growth to drive change. At the close of each qualitative interview participants were asked whether improving practice through performance evaluation or through professional growth had received more focus in their district over the past several years. As shown in Table 4.3 below, most participants relayed that professional growth had received more focus at their site (no patterns were noted by role, professional practice model, or rural status). It is important to note that the perceptions of the interview participants may not reflect the perceptions of the district as a whole.

Many of the participants elaborated on their responses, showing more depth in their choice. Answers that professional growth received more focus ranged from the practical to the philosophical. A teacher, for example, simply shared "professional growth—I don't think the evaluation has really done anything," while an evaluator at a different site posited:

> On the admin team [we] believe that professional growth is where the change is going to come from, it's not going to come through the evaluation system. We have to have one, it is what it is, just make it be the best it can be but almost let's not expect more from the evaluation system than we're likely to get from it.

Table 4.3. *Perceived Focus on Performance Evaluation or Professional Growth*

	Teacher(s)	Evaluator(s)	Supervisor
Site 1	PG	PE	Equal
Site 2	PG	PG	*n/a*
Site 3	Equal	PG	*n/a*
Site 4	Unsure / PG	Unsure	*n/a*
Site 5	PG	PG	*n/a*
Site 6	PE	PG	PE
Site 7	PG	PG / PG	*n/a*
Site 8	PG	Equal	*n/a*
Across Roles	PE (1), PG (5.5), Equal (1), Unsure (0.5)	PE (1), PG (5), Equal (1), Unsure (1)	PE (1), Equal (1)
Danielson	PE (1), PG (3),	Equal (1)	
Marshall	PG (1.5),	Equal (1), Unsure (1.5)	
Marzano	PE (2), PG (3)		
NBPTS	PG (3),	Equal (1)	
Rural	PE (1), PG (6),	Equal (2)	
Non-Rural	PE (2), PG (4.5),	Equal (1), Unsure (1.5)	
Across All Sites	PE (3), PG (10.5),	Equal (3), Unsure (1.5)	

Notes. PE = Performance Evaluation; PG = Professional Growth. In multi-site counts, where two interviews were conducted (e.g., two teachers at a site), each response was counted 0.5 to avoid over-weighting the site.

Answers that performance evaluation received more focus were largely rooted in the practicality of implementing the law:

> As much as the state rules make it look like growth is a piece, it really is an evaluation rule. If you read the actual law, it's very evaluation oriented as opposed to growth oriented and I think schools are just forced to deal with it that way.

Other respondents felt that performance evaluation and professional growth received equal focus, such as the teacher who shared: "We understand what's available to us to grow professionally. We understand what's expected of us to be effective. I can't honestly say that one is pushed more than the other."

The major changes underway at local school districts included shifts in the definition of effective teaching, cultural shifts away from teacher isolation, and an increase of focus on teacher goals and growth. The data show that these shifts yielded some benefits, some challenges, and some transitional difficulties. In the next section I describe findings of factors that participants shared as consistently beneficial.

Factors Perceived as Contributing to Increased Teacher Effectiveness

The second research question in this study was: "What factors do practitioners perceive as contributing to improved teacher effectiveness via teacher performance evaluation and/or teacher professional growth in their school or district?" Two major themes emerged in the data regarding factors practitioners perceive as contributing to improved teacher effectiveness: (1) the clarity of new standards and rubrics, and (2) efforts to develop "open-door" cultures with increased collaboration, trust, and transparency. These themes emerged through the qualitative data, the open response survey data, and the Likert scale constructs for performance evaluation

and professional growth (laid out fully in Tables 4.4 and 4.5 below with top responses highlighted in Figure 4.1 below).

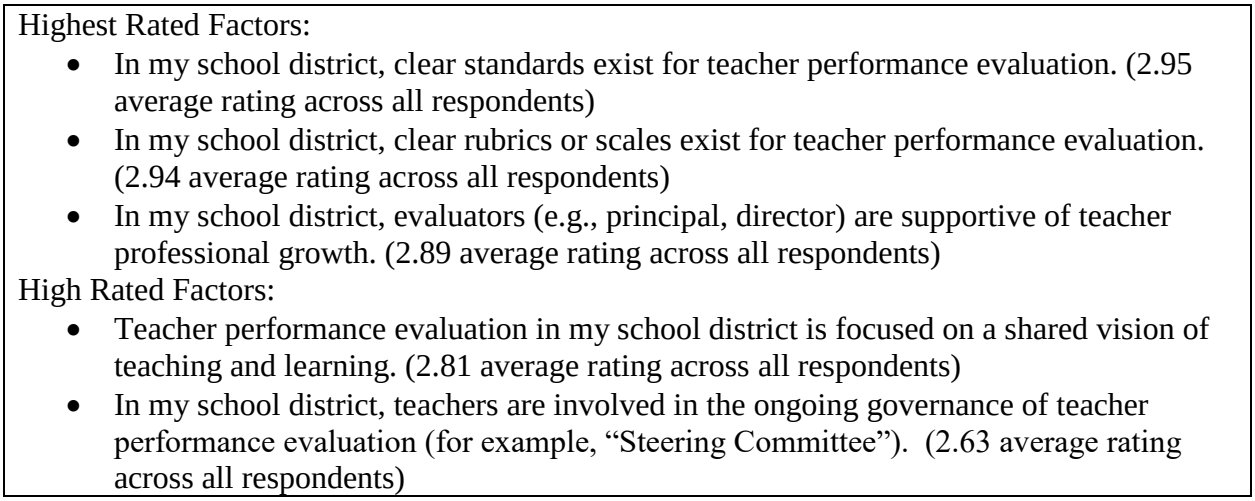

Figure 4.1. Factors contributing to increased teacher effectiveness.
Means are as follows: 1 = Strongly Disagree, 2 = Disagree, 3 = Agree, 4 = Strongly Agree.

The three factors labeled "Highest Rated" were each a top-four response at five sites in the survey data. The two factors labeled "High Rated" were less represented via within-case analysis: the "shared vision of teaching and learning" factor was a top-four response at three sites, and the "teachers involved in ongoing governance" factor was a top-four response at four sites but a bottom-four response at one site.

The only pattern noted by rural status with regards to the second research question was that respondents at rural sites described more teacher involvement in developing the process for performance evaluation. Interestingly, when comparing results by professional role, teachers were somewhat more likely than evaluators to share that performance evaluation is focused on a shared vision of teaching and learning, but somewhat less likely than evaluators to share that professional growth was based on such a shared vision or that the standards for teacher performance evaluation are relevant to teachers' responsibilities. Such a pattern was not noted in the qualitative data. One possible explanation for the distinction in the quantitative data may be

Table 4.4. *Performance Evaluation Factors Influencing Teacher Effectiveness*

	PE01 Teacher performance evaluation in my school district is focused on a shared vision of teaching and learning.	PE02 In my school district, the amount of teacher time needed for the teacher performance evaluation process is reasonable.	PE03 In my school district, the amount of evaluator time needed for the teacher performance evaluation process is reasonable.	PE04 In my school district, classroom observations are frequent enough for accurate teacher performance evaluation.	PE05 In my school district, clear standards exist for teacher performance evaluation.	PE06 In my school district, clear rubrics or scales exist for teacher performance evaluation.	PE07 In my school district, the standards for teacher performance evaluation are relevant to teachers' responsibilities.	PE08 In my school district, sufficient training is available for teachers to understand the process of teacher performance evaluation.	PE09 In my school district, sufficient training is available for evaluators to understand the process of teacher performance evaluation.	PE10 In my school district, the process of combining measures into a summative effectiveness rating is clear.	PE11 In my school district, the summative effectiveness rating accurately reflects teacher effectiveness.	PE12 In my school district, the summative effectiveness rating distinguishes between levels of teaching effectiveness.	PE13 In my school district, teachers were involved in developing the standards for teacher performance evaluation.	PE14 In my school district, teachers were involved in developing the process for teacher performance evaluation.	PE15 In my school district, teachers are involved in the ongoing governance of teacher performance evaluation (for example, "Steering
Danielson Rural	**3.27**	*2.73*	*2.64*	2.91	3.18	3.10	3.20	3.00	2.90	2.90	*2.80*	*2.80*	3.11	**3.22**	**3.22**
Danielson Non-Rural	2.89	2.83	2.66	*2.22*	**3.19**	**3.11**	**3.20**	2.91	2.91	*2.62*	*2.59*	2.68	2.75	2.94	**3.06**
Marshall Rural	**2.89**	2.82	**2.84**	2.45	2.76	2.78	2.75	*2.31*	2.46	2.47	2.50	2.47	2.67	2.58	*2.44*
Marshall Non-Rural	2.76	2.49	*2.23*	*2.02*	2.70	**2.84**	2.80	2.40	2.55	*2.34*	2.47	2.67	**3.05**	**3.05**	**2.81**
Marzano Rural	2.81	2.52	*2.10*	*2.13*	**2.94**	**2.97**	2.74	2.52	2.62	2.48	2.67	2.83	2.66	2.83	**2.93**
Marzano Non-Rural	**2.86**	2.53	2.50	*2.25*	**2.81**	2.66	2.54	2.37	2.41	2.29	*2.00*	2.31	*2.00*	*2.15*	2.32
NBPTS Rural	2.50	2.36	*2.15*	*2.00*	**3.07**	**3.07**	2.64	2.50	2.23	*2.14*	*1.69*	2.31	**2.75**	**2.67**	2.31
NBPTS Non-Rural	2.75	*2.06*	*1.97*	2.51	**3.07**	**3.03**	2.65	2.75	2.80	2.65	2.29	2.69	*2.07*	*2.14*	2.45
All Respondents	**2.81**	2.45	*2.31*	*2.31*	**2.95**	**2.94**	2.77	2.59	2.64	2.50	*2.37*	2.61	2.49	2.56	2.63
All Evaluators	3.09	2.59	*2.32*	2.86	**3.32**	**3.29**	**3.24**	2.86	2.90	2.95	*2.52*	3.05	3.00	3.14	3.05
All Supervisors	2.62	*2.08*	*2.00*	*2.23*	2.77	**3.08**	2.69	2.54	2.73	2.42	2.31	2.77	*2.09*	2.27	2.50
All Teachers	**2.80**	2.46	*2.33*	*2.27*	**2.92**	**2.90**	2.73	2.57	2.62	2.47	*2.36*	2.57	2.46	2.52	2.60

Notes. 1 = Strongly Disagree, 2 = Disagree, 3 = Agree, 4 = Strongly Agree. Note that the cross-site means (e.g., All Respondents, All Evaluators) over-represent sites with large n. Top-four responses in a row are bolded and aligned at the top of the cell; bottom-four responses are italicized and aligned at the bottom of the cell.

Table 4.5. *Professional Growth Factors Influencing Teacher Effectiveness*

	PG01 Teacher professional growth in my school district is focused on a shared vision of teaching and learning.	PG02 In my school district, collaborative teacher professional growth is well supported.	PG03 In my school district, individual teacher professional growth is well-supported.	PG04 In my school district, teacher professional growth is well-supported within the school schedule / calendar.	PG05 In my school district, teacher professional growth is well-supported outside of the school schedule / calendar.	PG06 In my school district, non-evaluative staff are available to help teachers grow (e.g., literacy coach, instructional coach).	PG07 In my school district, ongoing constructive feedback is provided to teachers.	PG08 In my school district, teachers are supported in reflective inquiry.	PG09 In my school district, evaluators (e.g., principal, director) are supportive of teacher professional growth.
Danielson Rural	3.00	2.89	3.00	*2.56*	2.89	**3.33**	2.89	2.89	**3.22**
Danielson Non-Rural	2.91	2.97	2.76	2.88	2.82	3.00	*2.41*	2.69	2.94
Marshall Rural	2.79	2.67	**2.82**	2.66	2.56	*2.06*	*2.36*	2.52	**2.97**
Marshall Non-Rural	2.56	2.54	2.61	2.44	2.50	2.51	*2.29*	2.41	2.73
Marzano Rural	2.80	2.80	2.80	2.56	2.68	*2.16*	*2.44*	2.60	**2.88**
Marzano Non-Rural	**2.79**	2.38	2.52	2.58	2.36	2.67	2.48	2.36	**2.76**
NBPTS Rural	2.33	2.33	2.58	2.42	**2.67**	2.25	2.17	2.25	2.58
NBPTS Non-Rural	2.66	**2.83**	2.81	2.55	2.60	2.38	2.57	2.55	**3.00**
All Respondents	2.72	2.70	2.73	2.58	2.60	2.48	*2.45*	2.52	**2.89**
All Evaluators	2.94	2.94	3.00	3.00	*2.59*	*2.50*	2.72	2.71	**3.17**
All Supervisors	2.54	2.83	**3.08**	**3.00**	2.77	2.77	2.85	2.85	**3.23**
All Teachers	2.71	2.68	2.69	2.53	2.59	2.46	*2.41*	2.49	**2.85**

Notes. 1 = Strongly Disagree, 2 = Disagree, 3 = Agree, 4 = Strongly Agree. Note that the cross-site means (e.g., All Respondents, All Evaluators) over-represent sites with large n. Top-four responses in a row are bolded and aligned at the top of the cell; bottom-four responses are italicized and aligned at the bottom of the cell.

that those in specialty roles (e.g., small-group interventionist, arts teacher) focused on the relevance of standards or growth activities to their atypical responsibilities. Alternatively, the distinction may be due to the more general nature of the "shared vision" prompt relative to the more tangible "standards relevant" prompt.

The Use of Standards and Rubrics to Support Increased Teacher Effectiveness

The benefits of increased clarity of professional practice standards and their rubrics were frequently raised by interview subjects, emerging in most of the interviews and as a major theme across interview subjects at three sites. Between the qualitative and quantitative data this clarity was consistently raised as a positive at nearly all of the sites, with few individual comments to the contrary. A supervisor expressed optimism about the new standards and rubrics at her site:

> I think it's going to be a really positive change and help us to look more carefully at everything that we're doing. I do think it will improve teacher effectiveness. In my own teaching and working with literacy coaches I am using the indicators more carefully. Before I didn't really rely on the ten teaching standards.

At this site and some others the district had shifted to a new professional practice model from Maine's previous teacher certification standards or other locally-developed standards.

In discussion about the benefits of the new, more clear standards and rubrics participants frequently described increased ability for teachers and the others to speak about teaching practice. One teacher shared, "I think what's changed is the ability for teachers as well as administrators to articulate what is effective teaching." This perspective was shared by an evaluator, who expressed:

> I think this whole process of changing our evaluation system has made us better in terms of being more mindful of the qualities that should be seen within a classroom, and what

teachers should be thinking about, and what principals should be seeing, and thinking about, and talking about with teachers.

This expanded articulation of effective teaching practice impacts growth as well as evaluation. With the level of detail expressed in the new rubrics, participants have ready access to a description of the next level of performance. At a different site, a teacher conveyed the utility of the standards and rubrics for personal goal-setting:

> With this system when I look at our rubrics I know where my areas are. This is really an introspective way of doing it that you have to self-evaluate. That's the first step in ours. I looked at it, and I looked at the highly effective, and even though I've taught for 27 years I don't feel like I'm highly effective in every area…are there things I can do better? Yes, I can see where I need to improve, and before I didn't get any of that.

Comments by evaluators in a group interview at a different site spoke to observing the rubrics initiate and support similar drive amongst their teachers. "When they're looking at those rubrics and they're only an effective teacher," one evaluator said, "there's that competitiveness in many of them to want to be highly effective….It shows them what they need to do, specifically, to move there."

A supervisor extended the perspective from self-assessment and internal motivation to overall benefit in the supervision, observation, and evaluation process. She shared that the specificity of the new standards and rubrics can ease tension for teachers while also contributing to growth:

> It's nice to know what [administrators] are looking [for] in regard to your classroom. It's no longer a mystery. You know exactly what they're looking for and you can showcase what you have and get feedback on what you need to improve on. And that's a shift you

need to make with staff, too, that it's okay to get feedback or commendations or recommendations as part of growing.

An evaluator using a different model shared similar thoughts about using the standards and rubrics to support the feedback process:

> I think the language in the rubrics helps to spark those conversations of what they are doing? What they aren't doing? What they could do a better job of? It's not just one word kind of section, but there's actual verbiage that you can start a conversation, or gather evidence for.

Overall, several supervisors and evaluators shared the benefits of increased ability for teachers to see the next steps for growth in an area, increased ability to give or receive actionable feedback, and increased ability to offer focused support to a teacher.

Some of those interviewed compared the teaching practice rubrics to rubrics for students in the era of learning standards and proficiency-based diplomas. For example, an evaluator at the high school level shared:

> They're not perfect, but I think they're good rubrics. I can see a progression from left to right, so I feel like they're really good feedback tools. If they were classroom rubrics, I feel like they'd be good rubrics for kids. I can point out, 'This is what I saw, and then the next level is this. It asks for you to demonstrate this, and I didn't see this, but do you think you did it or can you think of other classrooms that you've done it in?'

For that evaluator, the rubrics for teachers were consistent in approach to the school's efforts for students. An interview subject at another site said, "in the land of rubrics for proficiency and mastery with kids, now we're talking the talk with the adults too."

At multiple sites participants used the new, more clear standards and rubrics in conjunction with efforts to increase inter-rater reliability in teacher observation and evaluation. One evaluator, for example, shared how their site used a video-taped lesson of a teacher: "We sat down as a group and had conversation using the rubrics, to assess that teacher so that we're all on the same page." At some sites this calibration effort involved the Steering Committee, at other sites only evaluators participated in the inter-rater reliability work.

Overall, the new standards and rubrics were described as contributing to increased teacher effectiveness across the sites. A teacher emphasized the positive reception amongst colleagues of the new rubrics: "Even if the law went away, I think most of us would say, 'Even if you tried to simplify this we would still ask that you are evaluating teachers based on that new professional practice rubric.'" At many sites cultural efforts to improve effectiveness were similarly well-received.

Developing "Open-Door" Cultures to Improve Teacher Effectiveness

Factors related to culture were consistently raised by interview and survey participants in a positive way, though not as consistently as clarity of standards and rubrics. In the qualitative interviews when asked about successes in building teacher effectiveness participants repeatedly discussed efforts to build open-door cultures with more transparency, trust, and collaboration. Desired or realized outcomes of these efforts include increasing teacher and organization engagement with discussion of instructional practice, increasing teacher comfort with peers and administrators visiting their classrooms, and getting out of comfort zones in a supportive atmosphere.

At some of the sites efforts to increase transparency and trust were rooted in the development of and governance of the PE & PG system. At four sites the Likert scale data

showed relatively high ratings for teacher participation in the governance of performance evaluation (however this factor was rated as one of the lowest at another site). On another Likert scale indicator three sites had high-average responses regarding performance evaluation focused on a shared vision of teaching and learning (see Figure 4.1 above). A teacher at one such site shared her Steering Committee's approach to developing an aspect of the system, "As scary as it is for a lot of people, we've tried to make it as transparent as possible so people can voice their concerns." A teacher at a different site discussed the outcome of this shared approach:

> It's not a threat to them at all. It's not anything new…yes, we're checking all the boxes that the state has asked us to, but we've also made the system transparent K-12. There shouldn't be any question in anyone's mind about how teachers are evaluated by the administration… Every once in a while, you do have something unfortunate that happens, and administrators get brought in. Without a really transparent, open system people wonder, 'What happened to her?' or, 'What happened there?' I think the nice thing, this has built some trust, because it is so transparent. In general, I see a lessening of pressures and paranoia…There's not a gotcha here.

As quoted in the major changes section above, at a third site a supervisor spoke of the Steering Committee and administrative work behind making the standards and rubrics clear to teachers:

> Transparency and professional development around the look-fors… administrators have been doing a lot of unpacking of Marzano's framework and looking at what this would look like in a classroom at a four? At a three? Being transparent with teachers about that. These are the look-fors. Here's what we are hoping to find. The Steering Committee that I'm a part of, we do a lot of gathering of resources to help explain the components, the different design questions and elements.

Beyond trust and clarity, the involvement of teachers in the design, implementation, and governance of PE & PG systems had a positive side effect. At several sites consistent themes emerged in the interview data about incidental contributions to effectiveness for those who participated in PE & PG system development, such as learning more about standards, data, or instructional practices. A teacher shared, for example, "We watched the video, and then we each scored the teacher. That was eye-opening for the calibration piece. We were either a lot harder, or not even close in some cases to what they were scoring them using the rubric." At another site a teacher shared that she had learned much more about student learning standards through involvement in development of the SLOs.

Having worked to address inter-rater reliability through discussion and calibration, a principal and superintendent from the team of evaluators interviewed at a fourth site contributed this exchange:

> Superintendent: I think if nothing else, this process has made teachers feel more comfortable about administration being in the classroom and not feeling threatened by it.
>
> Principal: It just makes it transparent. I just had a conversation with six people on the rating. Here's the rubric, this is what I'm looking at…It's transparent…I think in the past it was based on one person's opinion. You didn't really have anything in writing.

However, at this and other sites evaluator and teacher perceptions may differ with regard to whether the benefit of classroom observations using more transparent criteria is yet realized.

A positive theme emerged in the qualitative data at three sites that conflicts with the quantitative data: perceptions were inconsistent regarding the benefit of mini-observations. At both Marshall sites and one Marzano site interview and open response data indicated benefit to

teacher effectiveness through a shift to more frequent and shorter observations (e.g., "walkthrough" or "mini" observations). For example, an evaluator shared that

> I think that when you stop by the classroom for three, four, or five minutes and watch what's going on everyday, or every other day, or however many you decide to do, you get a better picture of what's going on. I will tell you it changed tremendously in this building from teachers sitting at their desk to being up all the time and engaged with kids.

However, at two of those three sites the Likert scale responses for the prompt "classroom observations are frequent enough for accurate teacher performance evaluation" were in the bottom-four ratings for the site. This Likert scale item was the lowest overall response across all sites. Evaluators as a whole rated this item higher (2.86) than teachers (2.27) and supervisors (2.23). Further, at all three of those sites the Likert scale item "ongoing constructive feedback is provided to teachers" was a bottom-four response, and the fourth-lowest response across all sites. In this case, a difference existed but was not as drastic between the ratings of evaluators (2.72), supervisors (2.85), and teachers (2.41). Overall, in these data it was not yet clear as to whether the increased use of mini-observations was benefitting increased teacher effectiveness.

As part of encouraging open door cultures some sites deliberately sought to increase collaboration amongst teachers. This was discussed above with regard to a Professional Learning Communities approach to the SLO requirement, and appeared in the data in other ways. The theme of benefits to teacher effectiveness from collaborative efforts emerged in the qualitative interviews and open response data at four of the eight sites (e.g., collaborative professional development, collaborative planning work). The quantitative data did not conflict with this theme; however, the Likert scale prompt "collaborative teacher professional growth is well supported" was a top-four rating at only one site (average rating 2.70 across all

respondents). As noted in the data relevant to Research Question 1 (above) peer review was a major change underway at five of the eight sites; in generally early stages of implementation it emerged as a qualitative theme contributing to teacher effectiveness at two of those sites.

At the sites where the theme of building open-door cultures emerged, those interviewed shared that the transition to that culture was and is at times uncomfortable. As quoted in a previous section, one teacher supportive of her site's efforts realized it was taking colleagues out of their current comfort zone: "You have people asking you what you're doing, why you're doing it, which I think is a shift in general, in a positive way." Some teachers were comfortable with exposing their practice to more frequent visitors, while others were not yet comfortable with the increased scrutiny and more frequent affirming or critical feedback. A supervisor at another site discussed that the shift to more comprehensive and rigorous standards is part of that discomfort:

> One of the things I know that is tough for teachers is that they've been teaching for a long time ... If they're not marked as proficient, they get upset instead of recognizing that here are all the things that teachers can be held accountable for. But focusing on this area, you could use some room for growth. So that understanding that teaching is a lifelong learning process, too, I think, is hard for teachers.

The same supervisor shared how her role can contribute to increased teacher effectiveness, but that trust-building was also an important element for her:

> We need a liaison between staff and administration that can make things transparent, that can help teachers with their proficiency… it's a transition in that it's faculty are still learning to trust that it's okay to have someone come into your room that's not evaluative, and just there to help you with the process…There wasn't that transparency in our

> classroom where people could just come in and out of my classroom and see what I was doing. People would feel threatened by that. So that culture is changing.

An evaluator interviewed shared that in their PE & PG system design administrators must extend trust as well, in order for the peer review component to function:

> It's not monitored to do…it's a trust thing, like they're supposed to give each other professional feedback, there's no way for me to ... There's nothing I get, other than a little check saying they did it and so it's a trust factor. I think they're doing it though.

At that site and some others the researcher perceived that learning to trust and learning that gaps between one's current practice and high-level ratings was an ongoing process for staff in multiple roles. At some sites the development of transparency and trust were efforts multiple years in the making, while at other sites these efforts had begun in the months or year before the interviews, prompted by a recent change in personnel.

At sites where themes emerged about building "open-door" cultures the interviewer generally left the interviews with the perception that the development and implementation of the PE & PG system paralleled rather than prompted these cultural efforts. In some interviews the efforts to create "open-door" cultures were directly attributed to a change in school or district administration (e.g., Principal, Curriculum Director, Superintendent).

> The superintendent is a lot more in tune with the educators that are here and the one that was here for the two years prior ... came with a different toolbox. It didn't feel as if he really knew us as educators or where we were headed as a district.

When asked what prompted cultural efforts at their school district, a teacher interviewed at a different site simply replied with a district leader's name. The third-highest response in this quantitative construct indicated that evaluators were supportive of teacher professional growth;

this was a top Likert scale response at five of the six sites where culture emerged as a theme in the qualitative data.

Potential Rather than Realized Benefit

Some interview subjects (at three sites) had not yet seen benefits to increased teacher effectiveness under the new PE & PG systems, but were optimistic about the potential for such benefit. In particular, those interviewed were optimistic about the potential for differentiated or self-directed professional growth to improve teacher effectiveness. A teacher looked toward future years:

> I'm hoping once that money piece is gone it becomes less about, "Well, no I really did earn a three. You're wrong," and more about, "I earned a two there. Why did I earn this two? What can I do to improve? My evaluator is saying I could implement these classroom management strategies. Let's do it." I'm hoping the climate improves greatly as well once the money is gone.

An administrator also looked ahead to the potential of his district's PE & PG system:

> I think if the teachers buy in, honestly ... I like the model. I think it's maybe slightly overbuilt…I feel like there's a nice balance of administrative leverage as well as teacher input. It feels collaborative. The standards, like using the rubrics, they're not perfect, but I think they're good …I think if both parties are doing it…if it's grounded in the right stuff, I think it's a decent process. I think it's good.

These factors shared above were those perceived by participants as contributing to increased teacher effectiveness. Overall, participants were more inclined to share factors serving as barriers to that outcome.

Factors Perceived as Barriers to Improved Teacher Effectiveness

The third research question was: "What factors do practitioners perceive as barriers to improved teacher effectiveness via teacher performance evaluation and/or teacher professional growth in their school or district?" Three major such themes emerged in the data: (1) scarcity of time for staff, (2) transition challenges to the new PE & PG system, and (3) fundamental challenges that impede successful implementation and improved teacher effectiveness at some sites. These themes emerged through the qualitative data, the open response survey data, and the Likert scale constructs for performance evaluation and professional growth (laid out fully in Tables 4.4 and 4.5 above, with the lowest responses highlighted in Figure 4.2 below).

Lowest Rated Factors:

- In my school district, classroom observations are frequent enough for accurate teacher performance evaluation. (2.31 average rating across all respondents)
- In my school district, the amount of evaluator time needed for the teacher performance evaluation process is reasonable. (2.31 average rating across all respondents)
- In my school district, the summative effectiveness rating accurately reflects teacher effectiveness. (2.37 average rating across all respondents)
- In my school district, ongoing constructive feedback is provided to teachers. (2.45 average rating across all respondents)

Other Low-Rated Factor:

- In my school district, the amount of teacher time needed for the teacher performance evaluation process is reasonable. (2.45 average rating across all respondents)

Figure 4.2. Factors that are barriers to increased teacher effectiveness.
Means are as follows: 1 = Strongly Disagree, 2 = Disagree, 3 = Agree, 4 = Strongly Agree.

The factors above labeled "lowest rated" were each a bottom-four response at four or five sites in addition to a low overall average rating. The factor labeled "other low-rated" had a low overall average and was a bottom-four response at two sites. Rural and non-rural sites shared many of the same barriers, but additional barriers are described that largely emerged at rural sites. At all sites, regardless of rural status, scarcity of time appeared frequently in the data.

Scarcity of Time for Staff

Time for staff emerged as a robust thematic barrier to increased teacher effectiveness. This theme manifested in several ways: scarcity of time for evaluators (which impacts observation frequency), scarcity of time for teachers, and time and attention diverted from other efforts to the PE & PG system. These components are examined separately here, beginning with the most stated barrier.

Time scarcity for evaluators. In both the qualitative and quantitative data, time for evaluators emerged as the strongest factor serving as a barrier to improved teacher effectiveness and sustainability of the PE & PG systems. As shown in Figure 4.2 above, "the amount of evaluator time needed for the teacher performance evaluation process is reasonable" was tied for the lowest average rating, and was a bottom-four Likert scale result at five sites. Concerns about scarce time for evaluators directly emerged as a qualitative theme at all but one of the sites, and related themes (e.g., time scarcity for all involved, follow-through with planned system components) emerged at the eighth site.

In both the qualitative and quantitative data, evaluator time was raised as a major barrier by those in each sampled professional role. When expressing concerns for the time demand on evaluators, participants raised such topics as the high ratio of teachers to evaluators, the total number of planned observations and feedback conversations, and time-consuming documentation of observations and summative evaluations. An evaluator shared that "on the admin team, right now their biggest stressor is trying to figure out how to schedule it all…get into all the rooms…it didn't help that we had [15 new teachers] from last year to this year." A teacher elsewhere laid out a frank perspective on the time barriers for evaluators, especially in the context of recent budget cuts:

> I think it's tough because administrators are nervous and worried in how the heck they're going to pull this off…You're evaluating goals, you're doing observations, you're taking in the student growth... there's pre and post conferences, and…you're combining all this data for three years, you're coming up with a summative rating, you're tracking all of this, you're looking at artifacts, and our administrators, they're just trying to stay up with day to day stuff. They're going out of their minds… A lot of administrators have expressed that they can't even keep up with the pilot group, and that's not because of incompetence on the administrators' part, that's just the system is mind boggling…The problem is you can't have two administrators be around all the teachers all the time, handle all the discipline issues, all the parental issues, all the legal issues, all the school board meetings...At the high school level essentially it's a lot more work for those two administrators to handle. They're doing all the evaluation, all the discipline that used to be spread between three. That was done because of budgets.

At one site the Steering Committee attempted to reduce the time burden for evaluators by having trained peer observers share some of the load of observations and documentation; one such teacher empathized with the evaluators: "Being a peer observer, I would say it took me at least an hour to an hour and a half to just do the [observation] write-up. I can't imagine what it looks like from a summative point of view." At a different site an evaluator shared that some 20-minute observations can consume an additional two hours of documentation (this was in the context of following through with action plans for a teacher).

An evaluator at another site shared that time was greatly reducing his implementation of the planned mini-observation system for all teachers:

> The biggest challenge really is ... Okay, an observation. I have thirty-one classroom teachers, not to mention all of my specials teachers, my Reading Recovery teachers, my literacy coaches, so I have just there, trying to observe them and know what they're doing is an enormous task. Even getting in the classroom once a year doing a full observation, making sure I do all the paperwork for all those staff ... is a pretty large job given the amount of other stuff that I have to deal with… I really, honestly, don't imagine getting in ten times a year for every teacher, but my goal would be at least three or four. And teachers that have more needs than others, obviously will be more.

A teacher at that evaluator's school recognized and understood the cause of fewer mini-observations than intended:

> It's supposed to be a lot more than that. It's not happening, but I think they're just too busy… I think there's going to be a few unwieldy pieces of this and one of them is just getting in as often as you want to because of all the other things that you have to do. It's no different than most areas of teaching and education. It's just not enough time.

As noted above, a major change at school districts was an increase in frequency of teacher observation, shifting away from the clinical supervision paradigm to include more frequent observations at three-quarters of the sites, and to include mini-observations at nearly all of those sites. In the qualitative interview data, participants at only one site raised consistent concern that the new observation frequency was insufficient; that concern was primarily about sacrificing depth by using more mini-observations. However, in the quantitative data (see Figure 4.2 above) participants raised consistent concern, rating "classroom observations are frequent enough for accurate teacher performance evaluation" lowest among the Likert scale items. Furthermore, a low-rated Likert scale item was "ongoing constructive feedback is provided to

teachers." In the qualitative data this did not emerge directly as a consistent barrier to increased teacher effectiveness, though it was common for evaluator interview participants to express that it was challenging to complete the intended number of observations and feedback conversations:

> Ideally, at the end of the day do some sort of face-to-face conference…the reality of the day is that it's difficult to do. [Be]cause you have to meet with this parent…or you have to deal with this situation that just came up. And then it gets behind.

In related data, a theme emerged at one site that general challenges with follow-through of the intended system were a barrier to growth. The constant thread of evaluator time impacting observational frequency, ongoing feedback, and sustainability was the most major barrier to increased teacher effectiveness.

On the theme of insufficient observation frequency for accurate evaluation and ongoing constructive feedback, the qualitative and quantitative data may be showing differences because most sites were in a pilot stage of implementation. The evaluators, supervisors, and teachers interviewed were typically members of the Steering Committee or otherwise involved in PE & PG system design and implementation, so their perspectives as to frequency of observation, feedback, and accuracy of summative evaluation may reflect a system not yet experienced by a number of quantitative survey respondents.

Time scarcity for teachers. When discussing why one of the planned modes of peer review had not yet succeeded, a teacher replied that time was the limiting factor.

> There's never enough time and enough coverage. The amount of time to do all of it and do it well, and do it authentically, and have the discussion, teachers were using every single second of their planning period, lunch, anything they could find.

Time for teachers, either individually or in the context of supporting peer feedback, was raised in the qualitative data as a direct barrier to implementation and increased teacher effectiveness at three-quarters of the sites, and was indirectly raised at a seventh site. In the quantitative data, "the amount of teacher time needed for the performance evaluation process is reasonable" was a bottom-four result at only one-quarter of the sites, but was tied for the fourth-lowest overall rating across all respondents. This factor was not raised to the same extent as scarce time for evaluators, but still served as a barrier for increased teacher effectiveness, especially in the context of multiple other school, district, and state initiatives.

At half of the sites in this study, participants raised concern that time dedicated to the PE & PG system was negatively impacting teacher effectiveness by reducing school and teacher time spent on other beneficial efforts. A supervisor, who was generally positive about her district's implementation of the state law, shared:

> If you add more, something has to go…our time is really precious and [I'm] not always sure that focusing on some of these pieces is improving our student instruction. Does it always benefit students? That's our job. Sometimes I feel like we get distracted because we're trying to follow through elsewhere.

A teacher at the same site (also generally positive about her district's implementation) echoed those thoughts, explaining low teacher participation in a governance aspect of the PE&PG system:

> According to the plan, there is a committee that is going to be spawning off to do that piece, but there haven't been a lot of people jumping at the prospect…I would rather be spending time with kids. I think that's peoples' biggest problem, like that's the thing that stands in the way. People want to spend time with their students. They got into teaching

> to do that, not to fill out paperwork, not to be stacking their portfolio, they got into it to teach kids, and it's taking away from…teaching your students, which in turn makes you less effective. It's kind of a Catch-22.

Participants in other roles noted as well that teachers were reticent to leave their students, even when coverage was offered to reduce the need for substitute planning.

An evaluator at a different site raised concern with diverting attention from other efforts such as proficiency-based learning to implementation of the PE&PG system:

> I think it's those little things that are going to overshadow some of the really important things. Like some of the procedural things are going, at least for the next year or so. We're going to be so focused on those, and to me those aren't nearly so important, as the instructional piece it, and where I really want to focus on. But I think for the next year or so, that's what we're going to be focused on. And making sure everybody understands those things… I spent a lot of time stressing about…making sure that they have the time that they need, that they have the support. Because I don't feel right about telling somebody…to do things without the resources or the time to do it. It tears apart morale. It tears apart any business or school.

Similarly, an evaluator at yet another site shared "if we use up all of the capacity on learning iObservation, then that's like stealing attention that we might prefer was put on our staff book-read about project-based learning."

Time for evaluators, time for teachers, and time diverted from other initiatives connected as strong barriers to increased teacher effectiveness. Multiple years into the implementation process, an evaluator at another site was not optimistic about the sustainability of the overall PE & PG system:

> We have a great superintendent, we have great principals, and we are still struggling through this mess of trying to figure out what this system looks like. How on earth is this going to help teachers across the state of Maine and when will it get changed because people will find out how much work this is? The sustainability issue of, for instance, our high school principal has 50 teachers and he has one assistant principal who is helping him with evaluations, so each of them are having to evaluate 25 people twice a year and do all of those things, and then get kids into college, and worry about sports, and respond to parent calls, and fights in the hallway. I'm just worried with very good intentions the state is a requiring a system that is not sustainable, and so what's going to end up happening is a system that's just not carried out very well, that doesn't do what it was intended to do. Not because people don't care, but because there isn't time to do what's being asked well.

In addition to the time challenges raised above, other factors emerged as barriers to increased teacher effectiveness.

Transition Challenges and Complexities

As sites pilot and implement their new PE&PG systems, interview and survey participants raised a number of transition challenges that are currently barriers to increased teacher effectiveness; these challenges may be addressed over time as the new systems take hold and the intended cultures develop. As discussed above, some sites enacted cultural changes contributing to improved teacher effectiveness. Several sites, however, had strong barriers remaining in building staff comfort in opening doors and sharing professional practice. For example a supervisor shared,

> We wanted to put together a document where teachers could share their growth, where they felt they were strong as an educator. It just didn't work. We tried...That was for peer observations...A teacher brought it up that they would really like to know, because they aren't out of the classrooms a lot, who is really skilled at for instance assessments or their mini-lessons. We thought we would provide them teachers' names. When we tried to put it together nobody was willing or felt that they were good enough to be ... They thought that they would be judged. It's too bad, but understandable.

At the same site an interview participant shared that even though the Steering Committee is working hard to build culture and approach PE & PG with growth in the forefront, teachers still feel stress with what they perceive as high-stakes evaluation which makes it difficult for them to engage with the growth components.

At a different site, an evaluator shared difficulty in getting teachers to engage with the new PE & PG system.

> Their behavior speaks to me in a way that says that the evaluation system is something that is done to them. I would prefer that it be something that they're bringing to the table, and that we're having a professional dialogue, and I'm supportive of their professional growth…I want the teachers to come to that summative conference to present their evidence of meeting their standards. They often show up empty handed. They're waiting for me to tell them that they're an effective teacher as opposed to talking about what makes them an effective teacher…We need to come to the summative conference to talk about…this is how my students performed, this is what I've learned from it… It's interesting how many of them come, not really prepared, they'll rate themselves, but they'll, 'This is where I am.' It's not a really reflective conversation…

These factors currently serving as barriers to increased teacher effectiveness at several sites may be bumps in the road of initial implementation. Another large bump in the road at many sites is the complicated nature of their newly developed PE & PG systems.

Interview and survey participants repeatedly relayed large barriers to implementing and sustaining PE & PG systems that lead to increased teacher effectiveness: the complex and overwhelming nature of the new PE & PG systems and new standards for professional practice. This especially emerged through open response and interview data, but is also reflected in a bottom-four Likert scale response at three sites: "the process of combining measures into a summative effectiveness rating is clear." The researcher observed across the data collection that participants at several sites repeatedly looked up the PE & PG plans they were currently implementing to remind themselves of aspects of the system. Additionally, the researcher observed a number of self-corrections as to the weighting or timeline of components in the systems, and interview subjects who discovered new or missed aspects of their PE & PG system while referencing documents during the interviews. It was clear in several settings that the new PE & PG systems were sufficiently complex that even those deeply involved with the design and implementation of the systems had points of confusion or unfamiliarity.

Participants at both Marzano sites raised consistent concerns that the number of elements was overwhelming to teachers and administrators; proposals were underway at each site to re-focus on a smaller number of Marzano elements for at least the first several years. These efforts would sacrifice attention to the breadth of teachers' practice for the sake of manageability. A teacher shared:

> One of the challenges has been understanding the Marzano framework...It's overwhelming for some people. I think, just like our students, every teacher has a

> different learning style and for some people that list of things was perfect, "Oh, I know exactly what they're looking for"…And for other people it was an overwhelming amount of information to think about in terms of, "This is what they look for when they walk in my classroom, I can't even wrap my head around it."

The same teacher noted that "some really good veteran teachers left" because of the shift to the Marzano system: "They still probably had some years left to really offer the kids. Again, I think it was the sense that they had to jump through hoops they didn't want to jump through." An evaluator at the same site shared, "The number of elements is overwhelming to teachers who feel as though they are trying to 'cover' all elements at a surface level, rather than focus on pertinent elements in depth." Noting that not all elements may be observed or necessary in a particular lesson, that evaluator lamented "It feels a bit like 'one size fits all' taking away teachers' autonomy and creativity."

Issues of complexity were not limited to Marzano sites, however. A theme emerged at a site using a different model about the cumbersome nature of some system components; a teacher shared the concern that "the [paperwork] expectation is huge between before and after." At a site using a third model, an evaluator interviewed saw long-term issues of complexity in their PE & PG system design that may prove problematic in increasing teacher effectiveness:

> When push comes to shove and it comes to a disciplinary issue, there's so many loopholes, that we're kinda creating for ourself. You make one little mistake, and it could be grievable, it could throw everything out. And because that creates such a complex system…There's a lot of room for error. So I think a lot of administrators, to play on the safe side, may not do what they should do. And I think I'll be tempted to do that too.

At that site and one other, multiple interview participants relayed that the new PE & PG system is unlikely to help teachers who don't want to improve. Those participants noted that eventual removal of such teachers from the classroom would rely on pre-existing systems of action plans rather than the new PE & PG system.

Fundamental Challenges to Improved Teacher Effectiveness

Beyond challenges of time and the challenge of transitioning to a new and complex system, themes emerged about factors that serve as persistent barriers to improved teacher effectiveness. While not direct components of the PE & PG systems, these factors were strongly represented in the data at three of the four rural sites and influence successful implementation of PE & PG systems. Participants raised the following as barriers: staff turnover, geography, economy, and merit pay.

Turnover. Interview respondents and survey participants via open response at three of the four rural sites raised frequent administrative turnover as a major challenge at their school district impacting teacher performance evaluation and professional growth. A supervisor gave an example: "In my stay here, which is about probably 12, 13 years, we've probably had five or six superintendents…five or six principals. Three or four different assistant principals. At least two or three different special ed directors." At another site, an evaluator shared, "We've gone through six Superintendents in 12 years." The third such rural site at one point had nearly all administrators at the school and district levels depart in a single summer, leaving behind one administrator with only a short history in the district.

The turnover challenge was not limited to administrators. Interview participants at one of those rural sites shared that it was typical to have a new teacher for about two years before they moved south to wealthier districts. At the time of the interviews, months into the school year,

that site was still seeking to fill an open teaching position. At another rural site an evaluator shared a persistent recruitment barrier influenced by geography and funding resources: "We usually get applicants. Most of the time, they have the right certification. Pretty rare to find someone who's like, "Man, this is a great candidate.'" In contrast, at one of the non-rural sites the teacher interviewed shared background behind the district's approach to PE&PG:

> To be very honest with you, the evaluation process here in this district, we are very fortunate, we do not hire new teachers. We hire teachers that have an excellent track record. There are very few teachers that I would say are going to get a less than proficient or exceptional rating within their process.

Geography and economy. At two of the rural sites and one non-rural site lack of funding for professional development (especially teacher-driven professional development) emerged as themes in the data about barriers to improved teacher effectiveness. This professional development challenge was exacerbated at one of those rural sites by its relatively remote location; interview participants shared that most workshop, course, and conference opportunities for teachers were a lengthy drive away. Finally, at one rural site an interview participant shared that the funding barriers in the school district were shared by the overall community:

> In our area it feels like there's a lot less [employment] opportunities for parents and … families are hurting which impacts everything. Students are also struggling. We all are tied together. It's hard not to blame but also just keep doing the best we can. It's a little frustrating when it feels like the motivation to be a successful student has been lost.

Yielding and sustaining gains in teacher effectiveness face great barriers in those sites with these persistent challenges of staff turnover, geography, and funding.

Merit pay perceived as a barrier to effectiveness. As a final challenge to improved effectiveness, when the topic of merit pay arose it was viewed with a sense of skepticism (by a teacher) or in the negative (across four other sites by more than a quarter of the total interview sample). Some of those negative perspectives came from interview subjects who had direct experience with merit pay at their current district or a previous district, while others had colleagues who brought stories of divisiveness associated with merit pay.

For example, a supervisor shared:

> We did not want teachers to be paid extra for their scores, ever…I'm close friends with [someone in] another district that had done that and it created a lot of tension between teachers. That's the last thing we need. We need to be focused on students and we should be working collaboratively.

The teacher interviewed at the same site had similar thoughts: "With the philosophy that this is about bettering our entire school, and our community, and our population…money is definitely a motivator, but it also is a great divider."

An evaluator at a different district who'd experienced merit pay directly (as a teacher at a district not in the study) relayed,

> I don't think there's anybody that I know that enjoyed the merit pay side of it…It's a sham. I mean, it is, it is. I'd love to have the money, I got the money at the end, but the reason why I got the money was a farce. I just got lucky and that year my kids did well on the NWEA… It's just...you know, people gaming the system. Your group of kids. Your cohort of kids. How it works out that year. Yeah, sure, teachers have a direct impact on student performance. You know, but how that ties to a specific test on a certain day…I think it caused more harm than good.

An evaluator at a different site raised design considerations: “If Marzano intended this tool as a growth model, but then it's being used in a system where RIFing decisions are made, or pay increases are made based on your evaluation score, then it's no longer a growth model.” Raising ethical and motivational considerations, an evaluator with TIF experience shared:

> With an SLO it's really hard to know if all the information you're getting from a teacher is completely accurate. Not that most of our teachers would ever do this, but someone could very easily, I don't know, finagle the data or not do the process in the most trustworthy way. I think, unfortunately, part of that is because there is money at stake…If you're going to end up getting a two on your SLO rather than a three or a four that's the difference between thousands of dollars… I don't necessarily believe in incentive pay because I don't think it makes people do the work for the right reasons. I think people should do the work because that's our profession. You should be a teacher and a principal because you want students to achieve whether you get paid more money or not. You shouldn't work harder just because you're going to make more money. I think if anything it brings more ... I don't know. The only word I can think of is “yuckiness.”

Finally, a teacher who had personally financially benefitted from incentive pay at a TIF site shared:

> I will say that teachers have become very good at manipulating the system to get the score they want, to get the money they want. I'm curious to see what this will look like when the money leaves this year. I wish the money was not available… It's hard for me to perceive the scores that people are earning as valid when I see some of the behind the scenes stuff that teachers do to get those scores. Teachers always kind of put on a show when they knew their formal observation was going to occur. It's been stepped up quite a

> bit, that show. Teachers will talk now about what they're doing to get the best score possible. During a formal observation that's not something I see that carries over to day to day teaching… I think once the money has gone there'll be great potential…It's had a negative impact. I don't know if it's decreased our professional performance. Certainly our ethics have decreased… I hear people talk about the scores they earn in anticipation of the money they're going to get, not in reflection of the performance they exhibited and where they can go. I think since the grant has started I've had, as a teacher, a conversation with one teacher about their scores and what next steps they could take, and how they could improve, and I come into contact with a significant amount of teachers.

With the exception of the TIF district in this study, which was continuing merit pay for the remainder of the TIF grant, the researcher found no movement toward use of merit pay in this study, and typically passionate concern even amongst those who had benefitted or were likely to benefit from such a model.

Summary of Changes and Factors Impacting Teacher Effectiveness

In the preceding sections I described participants' perceptions regarding each of the research questions. Before launching into the final construct from the quantitative survey I summarize here the recurring themes. With regards to the first research question, numerous changes were underway at the sites in this study. Recurrent themes included: (a) new standards and rubrics; (b) the inclusion of student growth data in PE & PG systems; (c) expanding on the traditional observation model; (d) peer feedback; and (e) cultural efforts. Overall, interview participants perceived that their districts had placed greater focus on professional growth than on performance evaluation.

With regards to the second research question, participants perceived several factors contributing to improved teacher effectiveness. Major themes included: (a) the increased clarity of the new professional practice standards and rubrics, and (b) efforts to promote "open door" cultures with increased trust, transparency, and collaboration. With regards to the third research question, participants perceived several barriers to improved teacher effectiveness. Recurrent themes included: (a) time scarcity for staff, especially for evaluators; (b) challenges transitioning to a new and complex system; (c) persistent challenges with staff turnover; (d) challenges exacerbated by geography and economy (especially at rural sites); and (e) merit pay (where it was experienced). Some respondents discussed positive aspects of Student Learning Objectives, but most data reflected negative practitioner views about this mandate.

The data collection for this study, during the 2016–2017 school year, found one of the eight sites in its first pilot year for the new PE & PG system, five sites in the second pilot year, and two sites implementing their systems. Practitioners' perceptions may be different at later stages of implementation, especially as transitional challenges are addressed. In the next section I describe practitioners' early perceptions about the impact of their PE & PG system on teacher effectiveness.

Local PE & PG Systems Not Yet Impacting Teacher Effectiveness

When comparing the previous system of teacher growth and evaluation to the current system, there was not yet an overall perceived shift in teacher effectiveness expressed in either the qualitative or quantitative data. However, these data were collected at a quite early stage of implementation and qualitative interview participants were optimistic about the potential for improved teacher effectiveness. Notable differences on this topic were found in sub-sets of survey respondents. As described in the methodology, teacher effectiveness in this study is as

perceived and defined by the individual practitioners for the following reasons: (1) the term is not defined in Maine's PE & PG Systems regulation; (2) the local definitions, if they exist, would likely be unique to the professional practice model, locally-determined weighting of system components such as student growth data, and local values; and (3) even within the same school district educators' perspectives on effectiveness vary widely, drawing upon different logics and "intellectual, professional, and cultural histories" (Rigby, 2015, p. 378).

Teacher Effectiveness Perceptions Across the Eight Sites

The quantitative survey for this study used prompts to explore perceptions of differences in teacher effectiveness between the present day and prior to Maine's PE & PG law:

Q17. The following prompts are about your perceived changes since Maine's Performance Evaluation and Professional Growth law (PE & PG, 2012). For each, please select if you strongly disagree, disagree, agree, or strongly agree.

	Strongly Disagree	Disagree	Agree	Strongly Agree
Prior to Maine's PE & PG law (2012 and earlier), overall teacher effectiveness was strong in my school district.	○	○	○	○
Overall teacher effectiveness is currently strong in my school district.	○	○	○	○
Prior to Maine's PE & PG Law (2012 and earlier), my school district had a system of teacher performance evaluation that improved overall teacher effectiveness.	○	○	○	○
My school district currently has a system of teacher performance evaluation that improves overall teacher effectiveness.	○	○	○	○
Prior to Maine's PE & PG law (2012 and earlier), my school district had a system of teacher professional growth that improved overall teacher effectiveness.	○	○	○	○
My school district currently has a system of teacher professional growth that improves overall teacher effectiveness.	○	○	○	○

Figure 4.3. Perceptions of teacher effectiveness.
Likert scale is as follows: Strongly Disagree = 1, Disagree = 2, Agree = 3, Strongly Agree = 4.

Overall, the means on the above prompts, laid out in Tables 4.6 and 4.7 below, hovered between "Disagree" and "Agree." None of the means — for the whole group or any subgroup — reached the "Disagree" (2.00) level; the lowest mean of any group on a prompt was 2.18 for the current system of performance evaluation at the Marzano non-rural and NBPTS rural sites. A handful of subgroup means reached the "Agree" (3.00) level; six of these eight subgroups reaching 3.00 involved Danielson sites.

Table 4.6. *Perceptions Regarding Teacher Effectiveness (via Likert Scale)*

	Overall teacher effectiveness is strong		System of teacher performance evaluation improves overall teacher effectiveness		System of teacher professional growth improves overall teacher effectiveness	
Group	System Prior to Maine's PE&PG Law (2012)	Current System**	System Prior to Maine's PE&PG Law (2012)	Current System**	System Prior to Maine's PE&PG Law (2012)	Current System**
All Respondents (n = 239)	2.93 / .576	2.93 / .586	2.61 / .645	2.54 / .659	2.64 / .599	2.60 / .678
Danielson Rural (n = 9)	***2.78** / .441	***3.22** / .667	2.78 / .441	2.89 / .333	2.78 / .441	3.00 / .707
Danielson Non-Rural (n=29)	3.21 / .491	3.28 / .455	2.55 / .736	2.69 / .712	***2.59** / .682	***2.83** /.711
Marshall Rural (n = 32)	2.91 / .466	2.88 / .554	2.56 / .716	2.63 / .554	2.66 / .602	2.72 / .523
Marshall Non-Rural (n = 39)	2.87 / .570	2.90 / .552	2.54 / .600	2.56 / .552	2.59 / .549	2.59 / .549
Marzano Rural (n = 22)	3.05 / .575	3.09 / .610	2.64 / .492	2.68 / .568	2.64 / .492	2.73 / .631
Marzano Non-Rural (n = 28)	2.89 / .416	2.86 / .448	***2.68** / .476	***2.18** / .612	***2.75** / .441	***2.36** /.621
NBPTS Rural (n = 11)	2.64 / .924	2.36 / .809	2.55 / .688	2.18 / .603	2.55 / .522	2.27 / .647
NBPTS Non-Rural (n = 69)	2.91 / .636	2.86 / .576	2.64 / .727	2.54 / .759	2.62 / .709	2.51 / .779

Notes. Values are presented as mean / standard deviation. 1 = Strongly Disagree, 2 = Disagree, 3 = Agree, 4 = Strongly Agree. *Previous-current difference is significant at 0.05 level via paired samples t-test with listwise exclusion. **Differences between sites significant at 0.05 level exist via Kruskal-Wallis test.

As shown in Table 4.6 above ("Overall teacher effectiveness is strong"), for the set of all respondents the average rating was identical (2.93, with 3.00 representing "Agree") when comparing overall teacher effectiveness under the previous system to the current system. Four sites had lower average ratings under the current system than the previous system, and four sites had higher average ratings under the current system. Paired sample t-tests revealed that only the Danielson rural site showed a significant difference at the .05 level (rating effectiveness under the current system 0.44 higher than the previous system). For overall teacher effectiveness under the current system (but not the prior system) the Kruskal-Wallis test showed that a difference significant at the 0.05 level existed between sites. Post hoc analysis revealed numerous between-site differences significant at the same level: (a) between Danielson Rural and NBPTS Rural, (b) between Danielson Non-Rural and NBPTS Rural, (c) between Danielson Non-Rural and NBPTS Non-Rural, (d) between Marshall Rural and NBPTS Non-Rural, and (e) between Marzano Rural and NBPTS Rural.

The second pair of prompts focused on teacher performance evaluation and teacher effectiveness. As shown in Table 4.6 above, for all respondents there were overall decreases in means, rating the current system slightly lower than the previous system. The current system had a mean rating of 2.54, roughly halfway between "Disagree" (2.00) and "Agree" (3.00). Five sites had higher average ratings under the current system of performance evaluation, and three sites had lower ratings. Paired sample t-tests revealed that only the Marzano non-rural site had a significant difference within the site, dropping 0.50 points. For the current system of performance evaluation (but not the prior system) the Kruskal-Wallis test showed that a difference significant at the 0.05 level existed between sites, but post hoc analysis did not reveal a specific pair of sites that reached the 0.05 level.

The third pair of prompts focused on teacher professional growth and teacher effectiveness. As shown, for all respondents the current system was again rated slightly lower than the previous system, rating the current system of professional growth a mean of 2.60. Four sites had higher average ratings under the current system of professional growth, one site had the same rating under the current and previous systems, and three sites had lower average ratings under the current system. Across the eight sites paired sample t-tests revealed significant differences within two sites: the non-rural Danielson site rated the current system 0.24 higher than the previous system, and the non-rural Marzano site rated the current system 0.39 lower than the previous system. For the current system of professional growth (but not the prior system) the Kruskal-Wallis test showed that a difference significant at the 0.05 level existed between sites, but post hoc analysis did not reveal a specific pair of sites that reached the 0.05 level.

As a whole, these data presented in Table 4.6 above reveal few significant patterns when comparing the current systems of performance evaluation and professional growth to the previous systems. With the exception of one site, significant differences were not reported in overall teacher effectiveness, which should not be surprising at this very early stage of policy implementation. Similarly, with the exception of one site participants did not report a significant difference as to whether the prior or current systems of teacher performance evaluation improved teacher effectiveness. A teacher at a site not showing differences at the 0.05 level shared, "'Course, it's kind of early on, because only a handful of teachers have actually been through the new process. But at least [the Steering Committee is] getting feedback early on, which I think is good." Two of the eight sites showed significant differences as to whether the prior or current systems of professional growth improved teacher effectiveness, but they contrasted: one of those

sites showed a gain while the other showed a loss. The Marzano non-rural site had two sets of indicators with significant losses, but at this stage of implementation those data could reflect transitional challenges rather than overall system flaws. The teacher at that site shared, "I think of course one of the challenges has been understanding the Marzano framework. I think that's been a really big challenge. It's overwhelming for some people." That teacher had not yet heard the news the evaluator shared with me: "The teachers are not aware of this yet … [W]e are moving to the focused Marzano instead of the 66 elements, it will be… less than half of that…and I think they're going to love it. Because it's simplified, streamlined." Whatever led to the significant differences rating the current system lower than the prior system, either a more broad sample or more study at a later stage of implementation (e.g., three years into implementation) would help to determine whether those data were cause for concern.

Teacher Effectiveness Perceptions by Subgroups

As shown in Table 4.7 below, these means on the teacher effectiveness prompts were then compared by professional role, rural status, and professional practice model. Several trends emerged in those comparisons.

For overall teacher effectiveness under the prior system post hoc analysis following the Kruskal-Wallis test showed a significant difference between teacher and evaluator ratings. Evaluators rated the prior system a mean of 0.36 lower than teachers, but evaluators' ratings rose for the current system, closing much of that gap. Across all three pairs of prompts evaluators rated conditions under the current system higher than under the previous system (significant at the 0.05 level for "Overall teacher effectiveness is strong" via paired sample t-tests). However, while none were at significant levels, teachers rated the current system lower in each pair of

prompts, and supervisors rated the current system equal to or lower than the previous system in each pair of prompts.

Table 4.7. *Perceptions Regarding Teacher Effectiveness (Subgroups via Likert Scale)*

	Overall teacher effectiveness is strong		System of teacher performance evaluation improves overall teacher effectiveness		System of teacher professional growth improves overall teacher effectiveness	
Group	System Prior to Maine's PE&PG Law (2012)**	Current System**	System Prior to Maine's PE&PG Law (2012)	Current System	System Prior to Maine's PE&PG Law (2012)	Current System**
All Respondents (n = 239)	2.93 / .576	2.93 / .586	2.61 / .645	2.54 / .659	2.64 / .599	2.60 / .678
All Teachers (n = 210)	2.97 / .557	2.94 / .592	2.61 / .641	2.52 / .679	2.64 / .589	2.59 / .695
All Evaluators (n = 18)	***2.61** / .698	***2.83** / .618	2.39 / .608	2.72 / .461	2.44 / .616	2.67 / .594
All Supervisors (n = 11)	2.82 / .603	2.82 / .405	2.82 / .751	2.55 / .522	2.91 / .701	2.73 / .467
All Rural (n = 74)	2.89 / .587	2.91 / .666	2.61 / .615	2.61 / .569	2.65 / .535	2.69 / .618
All Non-Rural (n = 165)	2.95 / .572	2.94 / .549	2.61 / .660	2.51 / .695	2.63 / .627	2.56 / .701
All Danielson (n = 38)	***3.11** / .509	***3.26** / .503	2.61 / .679	2.74 / .644	***2.63** / .633	***2.87** /.704
All Marshall (n = 71)	2.89 / .522	2.89 / .549	2.55 / .650	2.59 / .550	2.62 / .570	2.65 / .537
All Marzano (n = 50)	2.96 / .493	2.96 / .533	***2.66** / .479	***2.40** / .639	2.70 / .463	2.52 / .646
All NBPTS (n = 80)	2.88 / .682	2.79 / .630	2.63 / .718	2.49 / .746	2.61 / .684	2.48 / .763

Notes. Values are presented as mean / standard deviation. 1 = Strongly Disagree, 2 = Disagree, 3 = Agree, 4 = Strongly Agree. *Previous-current difference is significant at 0.05 level via paired samples t-test with listwise exclusion. **Kruskal-Wallis test and post-hoc analysis revealed differences by role or professional practice model significant at the 0.05 level.

This trend did not stand out to the same extent in the qualitative data, though evaluators spoke frequently of potential under the new system: "So to me, that's going to be a huge ... I'm excited about it, and I think it has a lot of potential." At the forefront of the piloting or implementation process, evaluators generally seemed more comfortable than non-evaluators with the level of detail in the new, more specific professional standards and rubrics. Further, evaluators immersed in the new evaluation and growth system for many teachers have greater exposure to the impact or potential impact of multiple observations, goal setting, and in some cases of action plans. However, as shown earlier sustainability challenges for evaluators emerged in the data.

When comparing results by rural status a weaker quantitative trend emerged. While not significant at the 0.05 level via paired sample t-tests, the non-rural sites rated the current systems for performance evaluation and professional growth lower than the previous systems across each pair of prompts. In contrast, the rural sites gave equal or slightly higher ratings to the current systems. The Kruskal-Wallis test did not reveal any differences significant at the 0.05 level based on rural status.

When comparing results by professional practice model it was notable that participants rated the Danielson model higher under the current system on each pair of prompts. Paired sample t-tests revealed that these gains were significant at the 0.05 level for the first (plus 0.15) and third (plus 0.24) pairs of prompts. This upward trend held independently at the individual sites using Danielson (Sites 1 and 2). For overall teacher effectiveness under the current system, post hoc analysis following the Kruskal-Wallis test showed participants rating the Danielson model higher than both the Marshall and NBPTS models, significant at the 0.05 level. Kruskal-Wallis post hoc analysis also revealed that participants gave significantly higher ratings to the

current system of professional growth under the Danielson model than under the NBPTS model. It is important to remember the context of these data, however. Data collection captured the Danielson sites while piloting or phasing in their new PE & PG systems, thus many teachers were experiencing the systems for the first time. Further, teachers largely selected the Danielson model at each of those sites, which may influence the early data as the teachers conceivably feel more buy-in with the choice. Just as it would be too early to sound alarms for a site showing a decline across data points during the pilot phase, more study of these early perceived gains at Danielson sites would be appropriate (across a more broad sample and at a later stage of implementation).

Participants using the Marshall model were less consistent between sites but generally rated the current system equal to or higher than the previous system on the sets of prompts. Participants using the Marzano model did not show a consistent trend in these ratings; low ratings for the current system at Site 6 impacted the average between sites. Finally, the NBPTS model fared worst in these comparisons, with lower ratings for the current system on each pair of prompts at both NBPTS sites. These quantitative patterns were generally supported by the qualitative data, though in some instances the researcher perceived that an overall positive or negative level of satisfaction with the current administrator(s) perhaps colored both the qualitative and quantitative results.

These perceived shifts in teacher effectiveness, and whether the systems of performance evaluation and professional growth improve teacher effectiveness, must be taken in the context of pilot stage or early implementation for nearly all of the sites. As the changes made by districts are further implemented staff ratings for these prompts may well shift, and further study would be appropriate to capture a more mature stage of implementation. In the next Chapter I will

discuss these perceived changes in teacher effectiveness, the major changes enacted at districts, and the factors perceived to influence improved teacher effectiveness. In this discussion I will connect the findings to other contributions in the literature and to the conceptual and theoretical framework for this study. Finally, I will discuss the strengths and limitations of this study and provide recommendations for practitioners, policymakers, and researchers.

CHAPTER 5

DISCUSSION AND IMPLICATIONS

The purpose of this study was to examine perspectives from the field regarding major local changes in teacher performance evaluation (PE) and teacher professional growth (PG), the ways in which local PE & PG systems were or were not beginning to improve teacher effectiveness, and perceptions of factors contributing to or providing barriers to this improvement. In this mixed-method case study I collected and analyzed data from more than three hundred teachers, supervisors of teachers, and evaluators of teachers at eight sites (school districts) across the state. The data collection took place over the 2016–2017 school year, when each site was at the pilot or early-implementation stage of its PE & PG system developed to meet Maine's new laws and regulations.

The case study approach provides an in-depth look at the phenomenon and rich description, but is not intended to be generalizable. The breadth of data included in this study is considerable, and no other study in Maine has been found in a review of the literature to include both evaluator and non-evaluator perspectives in the wake of Maine's PE & PG systems regulations. The findings can inform the literature, the field, and policymakers about the varied approaches underway in Maine, supports and challenges that have arisen, and lessons learned by practitioners. The reader must understand that the findings should be viewed as informative and more study is warranted to gauge at later stages of implementation how well the policy is doing with its intended goals. In this final chapter of this dissertation, I briefly recapitulate the study before further discussing how the findings inform the literature and implications for researchers, policymakers, and practitioners.

Recapitulation of the Context and the Problem

In 2012 Maine passed the first of several laws and regulations mandating that school districts create teacher Performance Evaluation and Professional Growth systems (PE & PG systems) meeting particular criteria and shifting away from a long history of local control. Maine is a good example of such state reform in an era when federal pressures to change teacher evaluation and supervision increased via the *Race to the Top* program and *No Child Left Behind* flexibility waivers. One author noted that in this context states and districts "have been changing their policies at a dizzying rate, often with little chance to consider the research base for practice" (Darling-Hammond, 2013, vii). Hallinger, Heck, and Murphy (2014) reached a similar conclusion in their national research review described in Chapter 2, "the policy logic supporting [such] reform remains considerably stronger than the empirical evidence" (p. 5). In Maine, as in other states, a climate of change and unpredictability paralleled these efforts as the federal passage and implementation of the *Every Student Succeeds Act* removed much of the federal pressure to reform evaluation and supervision. Maine's PE & PG laws and regulation remain intact, though numerous revisions to relevant statute and regulation have been proposed or enacted by policymakers since the 2012 Educator Effectiveness Act.

In the piloting and early rollout of Maine's new PE & PG systems, an opportunity existed to conduct a study that addressed a problem affecting the field and the literature: little research existed nationwide about this policy shift that impacted many states, particularly in a context like Maine's where local districts retained some latitude in key elements of the system. Further, what research had been conducted was almost entirely from the perspective of evaluators rather than of supervisors or teachers (e.g., Goldring et al., 2015; Mason & Porter, 2015). There was a need, then, to provide further study that contributed to the literature and informed practitioners and

policymakers as they undertook implementation and considered refinement of these complex and important systems.

The stated purpose of Maine's changes is "to improve educator effectiveness by clearly setting forth expectations for professional practice and student learning and growth, and providing actionable feedback and support to help educators meet those expectations" (Maine, 2015, Section 1). Via the state mandates, policymakers combined the functions of formative supervision and summative evaluation. While the terms are at times misused in the field, evaluation and supervision are inherently different and lead to different results. Performance evaluation (i.e., summative evaluation) is primarily an administrative function to maintain minimum standards and determine employment consequences, while supervision is a formative function to provide direction and support for growth (Glickman et al., 2013). By combining these different functions Maine's mandates potentially exacerbate well-known tensions of role and authority when supervising and evaluating teachers (Oliva & Pawlas, 2004). Policymakers also enacted a controversial requirement to directly incorporate student growth data in teacher evaluation. The mandated changes in Maine came with few resources for school districts beyond one-time $4,600.00 grants; developing and implementing the requirements would largely draw from existing or reallocated local resources.

Prior to my research, the only studies of Maine PE & PG systems found in the literature were those conducted by the Maine Education Research Policy Institute (MEPRI). As described in Chapter 2, these MEPRI studies over four years included: multiple statewide surveys of more than 70 superintendents (Mason & Porter, 2014), interviews with seven superintendents or designees (Mason & Tu, 2015), interviews with school or district administrators in four districts in the initial PE & PG system piloting year (Mette & Fairman, 2016), and a statewide review of

PE & PG plans combined with administrator interviews in six districts the same year as my study (Fairman & Mette, 2017). Thus, data collection existed at multiple stages of local policy development and piloting, but the perspectives were limited to evaluators and written plans. My research builds upon the changes, successes, and challenges found by the MEPRI researchers by: (1) diversifying the data collection to include the perspectives of teachers and supervisors, (2) expanding the data collection to include the perspectives of many more practitioners, (3) expanding the number of sites studied, and (4) incorporating a mixed-methods approach at those sites. In addition to contributing to the literature I designed my study to inform policymakers about the early local effects of the law and to inform practitioners about successes and challenges realized or anticipated by those beginning to implement the requirements intended to improve teacher effectiveness.

Recapitulation of the Research Design

My study centered on three research questions examining the early local (school district) implementation of Maine's 2012 Educator Effectiveness Law, 2014 Performance Evaluation and Professional Growth Systems regulation, and 2015 revisions to law and regulation:

1. What do practitioners (teachers, supervisors, and evaluators) perceive as major changes in teacher performance evaluation and teacher professional growth in their school or district?
2. What factors do practitioners perceive as contributing to improved teacher effectiveness via teacher performance evaluation and/or teacher professional growth in their school or district?

3. What factors do practitioners perceive as barriers to improved teacher effectiveness via teacher performance evaluation and/or teacher professional growth in their school or district?

I used a mixed-methods multi-site case study approach to generate in-depth understanding of a complex situation (Creswell, 2008; McMillan & Schumacher, 2010) and to yield data on the present conditions and consequences of policies (McMillan & Schumacher, 2010).

After developing instruments, piloting the quantitative survey, and using multiple strategies to improve validity (McMillan & Schumacher, 2010) I recruited sites and collected data during the 2016–2017 school year. The sample includes eight sites: a pair of sites (one rural, one non-rural) using each of the four professional practice models typically selected by Maine school districts (Danielson, Marshall, Marzano, or the National Board for Professional Teaching Standards model). In addition to those primary site selection criteria sites ranged geographically over much of the state. I collected qualitative data via semi-structured interviews with a total of 20 teachers, supervisors of teachers, and evaluators of teachers (some sites did not employ non-evaluative supervisors). At each site I additionally collected quantitative data via an online survey for those in the same professional roles, adding 302 survey responses to the data pool.

I analyzed the data within-case and cross-case. In the first cycle for the qualitative data I used Attribute Coding, Evaluation Coding, and Values Coding (Saldaña, 2016), followed by second-cycle pattern coding to condense themes and concepts. For the qualitative data I used descriptive and basic inferential statistics (paired sample t-test, Kruskal-Wallis test with post-hoc analysis) to describe practitioner perceptions and determine patterns in those perceptions across professional role, professional practice model, or rural status. These efforts resulted in a

comprehensive study design for the intended purposes; as in any study strengths and limitations should be explicitly stated.

Limitations, Strengths, and Trustworthiness

The study design yields limitations and strengths; I describe these here and briefly recapitulate my efforts to maximize trustworthiness (these efforts are described in detail in Chapter 3). First, of the 164 school districts across the state of Maine, the study involves a limited sample of eight districts, including 20 interviews and 302 survey responses. The case study approach provides an in-depth look and rich description at those sites, but cannot be generalized to the whole state.

The study is slightly limited in its coverage of the state; approximately 14% of school districts in Maine were eliminated from site recruitment. These sites either: (a) did not choose one of the four professional practice models in the study, (b) had not yet received initial state approval of their PE & PG system plans, or (c) were in my immediate professional network in the Penquis region. The largest limitation from these recruitment choices is the study will not reflect the perspectives of districts that chose a professional practice approach other than the models pre-approved by the state (Maine's Chapter 180 allows alternative approaches that meet certain criteria).

Next, participation at the site level after recruitment contact was determined by a district gatekeeper. In the 15 failed recruitment efforts typical replies were either: (a) no response, (b) a response that the study did not fit at the present time, (c) a response declining to add more to teachers' plates via the survey, or (d) a response that the district or administrators had too much going on at the moment to participate. It would be excessively speculative to guess at what lies behind non-responses or responses regarding fit. However, response types declining to

participate due to time concerns indicate that the major finding of time barriers spreads beyond the districts in the study.

Putting energies into casting a wider net to include more sites than previous studies in Maine means hearing less from each included district. At each site, there were a small number (2–3) of people interviewed; others may have had different views. The overall survey response (302 participants, approximately a 33% overall response rate but ranging at sites from 15% response to 66% response) includes more participants than previous studies, but is still limited. At each site we cannot know if the views of respondents correspond to the views of those who did not respond, or what motived individuals to respond or not respond. The survey data are also self-reported data, which are more subjective by their nature.

A final limitation in this study is that the qualitative data collected are from participants identified by the district gatekeeper rather than randomly selected teachers, supervisors, or evaluators. Thus, it is possible those participants are more likely than their peers to hold views that the gatekeepers wished the researcher to see about their school districts or about PE & PG systems. This final limitation is partially alleviated by the mixed-method nature of this study; all teachers, evaluators of teachers, and supervisors had an opportunity to participate via the survey which included checklist, scaled, and open-response items.

This study has strengths beyond the triangulation enabled by the mixed-method design. The inclusion of three professional roles (teachers, supervisors of teachers, and evaluators of teachers) provides a multi-faceted perspective about the changes underway in Maine and the factors contributing to or arising as barriers to increased teacher effectiveness. The study of the professional practice models used by 93% of Maine districts (Lomonte, 2015) and the study of both rural and non-rural sites for each model increases the likelihood that the results represent

what is experienced elsewhere in Maine. This perspective is similarly broadened by use of a sample that covers a wide range of district population, of financial resources, and of Maine's geography. Further, the four models in the study are each nationally available, so the results may be informative to those outside of Maine.

Finally, I increased the trustworthiness of this research by proactively identifying and developing a plan through which to bracket my biases (see Chapter Three). This plan was particularly reflected through my use of researcher notes to identify my reactions to site visits and choices in data analysis, and through my repeated returns to the raw data to ensure accurate analysis. Trustworthiness was further enhanced by my consultation throughout with Dr. Ian Mette, my advisor who has conducted extensive peer-reviewed qualitative and quantitative research on related topics.

Summary of Major Results and Observations

This study yielded results addressing all three research questions: the school districts in this study have made major changes in the wake of Maine's Performance Evaluation and Professional Growth Systems laws and regulations, and factors contributing to and providing barriers to improved teacher effectiveness were identified. Changes underway at the sites included: (a) new professional practice standards and rubrics; (b) the inclusion of student growth data in PE & PG systems; (c) expanding on the traditional observation model; (d) peer feedback; and (e) cultural efforts. Most of those interviewed perceived that their districts had placed more focus on professional growth than on performance evaluation. Interview and survey participants shared several factors that they perceived contributed to improved teacher effectiveness: (a) the increased clarity of the new standards and rubrics, and (b) efforts to promote "open door" cultures with increased trust, transparency, and collaboration. Participants perceived several

barriers to improved teacher effectiveness: (a) time scarcity for staff, especially for evaluators; (b) transitional challenges as they moved to a new and complex system; (c) persistent challenges with staff turnover; (d) challenges of geography and economy (especially at rural sites); and (e) merit pay (where it was experienced). Some respondents discussed positive aspects of Student Learning Objectives (e.g., using SLOs for collaborative data work), but most data reflected negative practitioner views about this mandate: time-consuming, "scary," "a game that anyone can win," "they cause way more problems than they're worth."

One of the problems leading to this study was that the perspectives of evaluators (e.g., principals) dominated the small amount of literature on these new evaluation and growth systems. In my research design I deliberately sought the perspectives of teachers and non-evaluative supervisors. Most districts in my sample did not employ non-evaluative supervisors, but there were differences found in the views of teachers and evaluators, primarily through the quantitative data. Some item-by-item differences did not illuminate consistent trends. It was clear, though, that teachers disagreed with evaluators as to: whether observation frequency was sufficient for accurate performance evaluation, whether ongoing constructive feedback was frequent enough, and whether overall teacher effectiveness was improving. The data showed far more topics on which teachers' and evaluators' perspectives were in alignment, even in areas where one might anticipate differences would emerge. For example, those in both roles shared discomfort with the prospect of merit pay, distaste for the Student Learning Objectives requirement, and agreement that the local focus had been on growth over evaluation. Teachers raised as much concern as evaluators regarding the time demands on evaluators.

In this study I also deliberately sought perspectives from rural and non-rural sites. At rural sites, the data showed more teacher involvement in developing the process for performance

evaluation than at non-rural sites. Most of the distinctions based on rural status, though, were barriers that challenge rural school districts in implementing and sustaining this large change to evaluation and supervision: a smaller number of staff who must wear many hats, overall funding, staff recruitment, staff retention, geographic distances to professional development, and rapid administrative turnover.

Quantitative prompts explored perceptions before and after the state mandate regarding effectiveness, performance evaluation, and professional growth. At this piloting or early stage of implementation respondents did not yet perceive an overall shift in teacher effectiveness. Some statistically significant results existed on the prompts via paired-sample t-tests with listwise exclusion and the Kruskal-Wallis test with post-hoc analysis. However, with districts piloting or in early implementation at the time of study this is a very early stage at which to explore the outcomes of policy. Such quantitative trends may be highly colored by such things as the sample size, early transition challenges, the approval process, or changes in leadership. More study with a larger sample and at a later stage of implementation would be appropriate before drawing conclusions about whether the policy is meeting its intended goal. Below I discuss the findings in the context of the literature, consider likely and alternative explanations, and identify remaining gaps in need of future research. I conclude the chapter by summarizing the implications of my study for researchers, policymakers, and practitioners.

Discussion

To frame the discussion I return to the theoretical and conceptual framework for this study, in which school districts navigate an evaluation-supervision tension. This tension is illustrated in Figure 5.1 below (reprinted from Chapter Two). The well-regarded "SuperVision for Successful Schools" model (Glickman, Gordon, and Ross-Gordon, 2013) is representative of

much of the literature, focused on the development and growth of teachers through supervision. Maine's new law and regulation (expressed at the most detailed level through Chapter 180) include elements of supervision but are textually more focused on the measurement and evaluation of teachers, a policy logic described by Hallinger, Heck, and Murphy (2014). As portrayed in Chapter Two, while there are overlapping features the law and the literature differ in approach and local school district implementation is likely to include major elements of each, without fully encompassing either approach. Finally, school district implementation is inclusive of local assets and values and limited by local constraints.

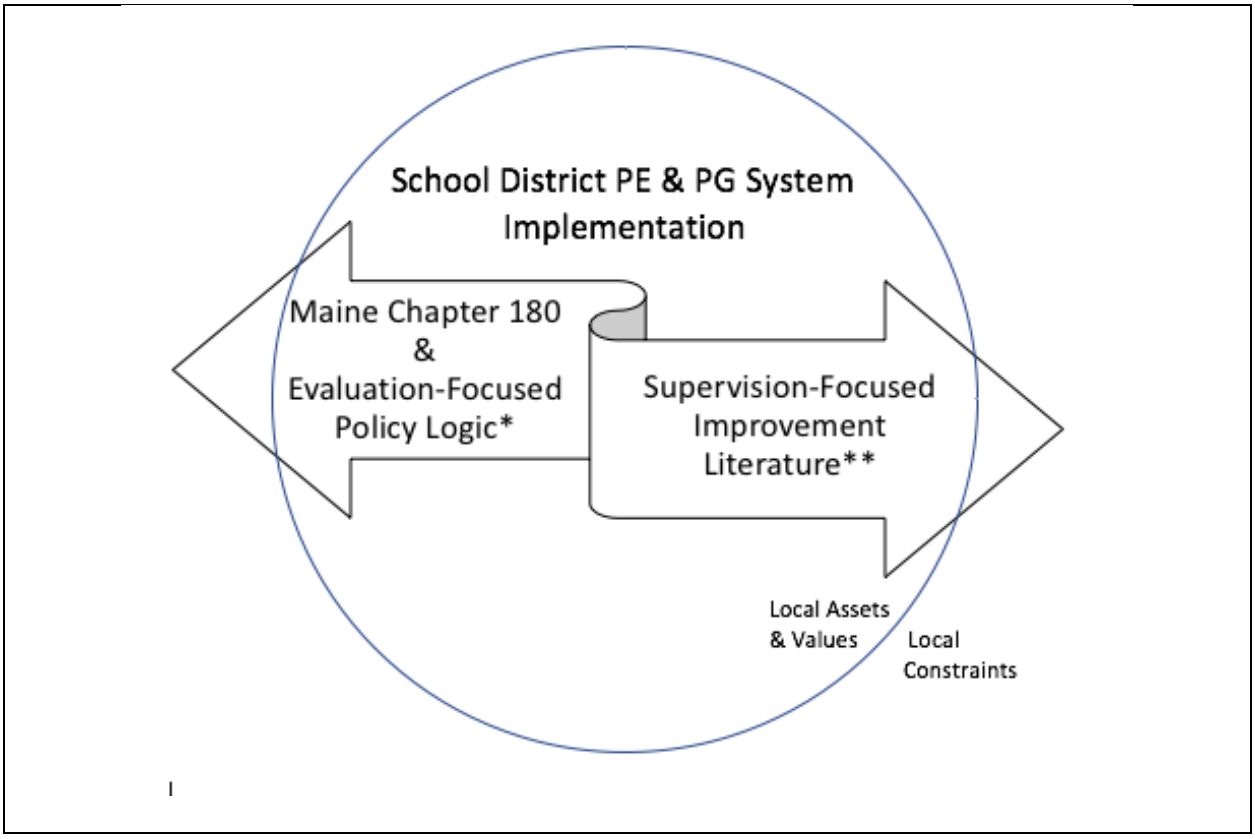

Figure 5.1. Conceptual framework for this study (reprinted from Chapter 2).
*e.g., as described by Hallinger, Heck, & Murphy, 2014
**As represented by SuperVision for Successful Schools (Glickman, Gordon, & Ross-Gordon, 2013)

We need not delve deeply into the status of districts relative to law and regulation; each of the sites in this study had received at least initial state approval of their PE & PG system and

was either piloting or fully implementing their plan. Thus, by the state's determination each site's plan complies with Chapter 180 (though it is notable that one site delayed piloting the student growth data requirement in the hopes that the requirement would be removed from law before implementation). In discussion then I will focus on the sites' location relative to Glickman et al.'s "SuperVision," navigating the tension between supervision and evaluation.

As described in Chapter Two, Glickman et al. (2013) assert that successful schools must move from a *conventional* or *congenial* model to a *collegial* model for success; I briefly remind the reader here of the major features of each. According to Glickman et al., the *conventional* model (which results in dependency, hierarchy, and professional isolation) is characterized by a focus on inspection and attempts to control teacher behavior; Maine's evaluation-focused mandate could be implemented in a way that largely results in the *conventional* model. The *congenial* model (which results in ineffective practices, inefficient use of time, and professional isolation) is recognized by friendly social interactions and a culture in which teachers largely have the latitude to perform as they please. The *collegial* model (the goal of Glickman et al.) is distinguished by, in summary: (1) collaboratively developing and implementing a vision for teaching and learning; (2) purposeful growth-focused (rather than compliance-focused) collaborative adult interactions improving schoolwide teaching and learning; (3) quality instructional supervision in a way that minimizes hierarchy and maximizes collegiality; and (4) deliberate development of knowledge, interpersonal skills, and technical skills to support these efforts. In my study the qualitative and quantitative data collected from over 300 practitioners at eight sites revealed where school districts were focusing their efforts as they implement Maine's mandates, navigate the evaluation-supervision tension, and attempt to improve teacher effectiveness. I now discuss the major findings (summarized above) in the context of the

research questions, the literature as a whole, and the conceptual and theoretical framework for this study.

Attempting to Improve Teacher Effectiveness Through Policy-Driven Changes

A major portion of changes underway at local school districts were driven by the new law and regulation. These included governance of PE & PG systems, adopting certain professional practice standards and rubrics, using more evidence-based processes, developing inter-rater reliability (e.g., Bradshaw, 2002; DeSander, 2000), and the use of Student Learning Objectives in PE & PG systems. I discuss those changes here.

Governance a shared vision of teaching and learning? In developing a PE & PG system plan each site (under state regulation) formed a Steering Committee that involved teachers and administrators, bringing multiple stakeholders into the process of selecting a professional practice model and choosing or developing system components. This approach on its face meets the goal of Glickman et al. (2013) of a collaboratively developed vision for teaching and learning, however the depth of the collaborative vision work was limited. Visions of teaching and learning not congruent with the policymakers' direction as expressed in law and regulation would not be approved by the state. Thus at minimum there was an element of hierarchy, with strong state-driven parameters on the vision. Supporting this notion, Fairman and Mette (2017) found that more than 95% of Maine districts had chosen one of the professional practice models for teachers pre-approved by the state.

At one site, the selection of professional practice model was largely delegated to teachers, resulting in a model that was mutually acceptable but not truly collaborative. At some sites in this study the collaborative vision-making was likely stronger; a few sites chose locally to implement one of the four professional practice models before the state passed laws and

regulations mandating their use. However at most sites the selection process was driven by either the state mandate or seeking Teacher Incentive Funds.

Ultimately, to whatever degree policymakers' influence inhibited Glickman et al.'s (2013) *collegial* vision-making, the participants in the study using the Danielson, Marshall, or Marzano models were at least neutral if not satisfied that the result was a shared vision. Shared vision indicators at the NBPTS sites were neither a strength nor one of the greatest barriers to increased teacher effectiveness. In qualitative data it emerged that those sites had chosen the NBPTS model under either financial or time pressures. As described earlier the NBPTS sites in my study had the weakest results for perceptions of improved teacher effectiveness (rating effectiveness higher under the previous system than the current system). These results may be associated with the shared vision indicators; if practitioners at the site have not developed agreement about what good teaching and learning looks like, whatever efforts they make toward improved effectiveness will not likely be in concert (e.g., Stronge & Tucker, 1999).

Across all models and all employment roles the quantitative responses regarding a shared vision of teaching and learning were stronger for performance evaluation than for professional growth, despite practitioners sharing in interviews that growth is more in the forefront of their efforts. Another potential weakness emerged within the quantitative data, perhaps indicating that collaboration across stakeholders may not have been present for some important steps. Specifically, the general responses for a shared vision of teaching and learning were stronger than when broken down as to teacher involvement in developing the standards, in developing the process, or in ongoing governance for performance evaluation. There are several possible explanations for these differences. First, evaluation may have been in the forefront of decision-making when selecting models and developing the PE & PG process. Hazi and Ricinski (2009)

and Ponticell and Zepeda (2004) note that over time evaluation has dominated supervision in educators' minds: "for all teachers and for the vast majority of principals supervision was, quite simply, evaluation." (p. 47). Second, numerous participants interviewed described elements of teacher saturation or fatigue in involvement (preferring to focus on their students), or on shifts in attention to other initiatives such as proficiency-based learning. As districts shifted from initial development of PE & PG systems to piloting or implementation, and as other initiatives emerged, the sense of shared involvement may have begun to decrease.

Going forward these findings indicate that a potential behind-the-scenes weakness exists that may inhibit reaching or sustaining the desired *collegial* model. The shared vision of teaching and learning is the foundation for much other work, so work to reinforce and sustain that vision is warranted. At most sites this appears to be a moderate weakness. However, at the NBPTS sites in this study, where the model selection was driven by financial or time pressures, it may be difficult to build upon the relatively weak shared local vision for teaching and learning.

Beginning to implement the shared vision through standards, rubrics, and common understanding. With Maine districts having partially achieved the collaborative vision-making needed for Glickman et al.'s *collegial* schools, attention then turns to the detailed standards and rubrics of those models created by national organizations and selected by local school districts. Participants' views on the standards and rubrics were generally quite positive, noting the ability to articulate effective teaching, to use the rubrics as a personal growth tool, and to use the rubrics to support a teacher in their growth. These are outcomes consistent with the literature on standards in systems of supervision and evaluation (e.g., Holtzapple, 2003), and with the emerging research in Maine (Mette & Fairman, 2016; Fairman & Mette, 2017).

The participants raised fewer concerns than strengths about the new standards and rubrics. The concerns raised were generally about the overwhelming nature of demonstrating and documenting the many facets of effective teaching, the practicality of observing and evaluating each element, or, by a few participants, about the unintended consequence of teachers sometimes narrowly tailoring instruction to satisfy rubric needs instead of student needs. This last consequence—perceiving that each moment of observed instruction needs to reflect every element of a rubric or checklist—has been previously found in the literature (Darling-Hammond, 2013). With these new professional practice models that concern may be a transitional challenge that fades as teachers and those visiting their classrooms gain familiarity with the rubric elements and especially if evaluators meet their aspiration of putting growth purposes rather than fear-inducing evaluation purposes in the forefront. In research conducted during the same school year as this study, Fairman and Mette (2017) found that sites were making progress in understanding the new professional practice models and yielding more support for professional growth.

The challenges for teachers and evaluators of demonstrating, observing, evaluating, and documenting each rubric element are less likely to fade in implementation and therefore structural adjustments may be warranted; such time challenges were findings in both of the most recent MEPRI studies as well (Mette & Fairman, 2016; Fairman & Mette, 2017). Both Marzano sites in this study were in the process of re-focusing on a smaller number of elements for the sake of manageability (Marzano's 2014 model includes 60 elements across four domains; the 2017 model maintains four domains but focuses on 23 elements). To put that in perspective an evaluator with a caseload of 25 teachers would have 1500 elements to assess each summative evaluation cycle under the 2014 Marzano model, or 575 elements to assess each summative

cycle under the 2017 model. Under multiple models participants shared promising emerging practices such as enabling teachers to provide "off-stage" evidence of practice (e.g.; Marzano & Toth, 2013; Stronge & Tucker, 1999) to their evaluators or for peer observers to confirm the use of practices that were not seen during an evaluative observation.

Maine's regulation requires training for evaluators and teachers; Maine evaluators recognize the need for ongoing calibration (Mette & Fairman, 2016; Fairman & Mette, 2017). For evaluators efforts to build inter-rater reliability (e.g., Platt et al., 2000) were common in the study, using strategies such as video-taped lessons (Glickman et al., 2013) or instructional rounds toward this end. These efforts address reliability challenges noted by many authors in the literature (e.g., Platt et al., 2000), and potentially impact the tendency for administrators to inaccurately assess poor performance (Weisberg et al., 2009). Maine's regulation is more vague about training needed for teachers on the standards and rubrics and the practices at districts were also less consistent. At some sites teachers were directly involved in the inter-rater reliability efforts, or the Steering Committee was putting significant effort into providing examples of varied level of practice on the standards and rubrics. At other districts teachers were less involved in such activities, which may result in the lack of common technical language for discussing teaching practice that has inhibited previous systems (Spillane, Reiser, & Reimer, 2002).

Overall these detailed standards and rubrics provide a strong tool practitioners can use to implement the *collegial* vision, with the potential to be used primarily in a growth-focused rather than compliance-focused way. The developmental nature of the rubrics can contribute to other elements called for by Glickman et al. (2013): aiding quality instructional supervision by providing a common language and through the developmental rubrics, and aiding the deliberate

development of knowledge and skill. If school districts are able to address the challenges of implementing and sustaining these complex but helpful professional practice models, there is promise for significant long-term payoff in teacher supervision and evaluation.

The use of student learning objectives. Some authors in the literature promote the use of student learning data in teacher evaluation systems (e.g., Wilkerson, 2000); the use in Maine was mandated by the state under federal pressure. This was a major change at each of the sites in the study, and each was implementing use of Student Learning Objectives (SLOs) due to the state mandate rather than due to local initiative. Where positive aspects were raised, participants spoke of a growth rather than evaluative focus, engaging teachers in goal-setting and individual or collaborative study of the data which some referred to as a Professional Learning Communities approach (Dufour & Eaker, 1998). Negative perspectives regarding SLOs were much more frequent, with participants invoking fear, challenges for those who serve small numbers of students, challenges for those in infrequently-tested subjects, validity challenges, and consumption of large amounts of time for little perceived benefit.

Similar difficulties were found by other researchers. In a 2015 Maine Educational Policy Research Institute report Mason and Tu interviewed seven Maine superintendents "to assess issues involving the incorporation of student growth data in their PE/PG system" (p. iii). Amongst that sample of district leaders implementing student growth data in evaluation earlier than strictly mandated, they found many of the same challenges practitioners raised on that topic in my study. Mason and Tu summarized some of their findings: "Researching and selecting specific instruments often involved a complex and time consuming process conducted by teams of faculty and administrators" (2015, p. iv). Continued challenges with the use of student growth

data arose in each of the most recent MEPRI reports (Mette & Fairman, 2016; Fairman & Mette, 2017).

To achieve Glickman et al.'s *collegial* schools, the use and analysis of assessments and data to inform instruction would likely be an important part of any vision for teaching and learning and of deliberately developing practitioner knowledge and skill. This view is reinforced by other authors (e.g., Darling-Hammond, 2013; Marzano & Toth, 2013). Participants in my study did not voice concern with assessments and data in general. However, across models, across roles, and in most of the districts in my study inclusion of SLOs into the PE & PG system was compliance-focused, time-intensive, and widely viewed as unhelpful toward improving teacher effectiveness.

I recognize that practitioners (including me) are bound by state mandate to continue implementing SLOs, and also recognize my own preference to focus on the analysis of data rather than raw student growth data in evaluation (a preference I've held as both a teacher and as an administrator). I view analysis of assessment data to inform curriculum and instruction as tremendously important. I also view analysis of assessments as a piece of a complex puzzle, and support our educators in combining diverse data points with their professional judgment to make educated guesses about the right curricular and instructional moves for students. However, challenges emerge the moment the outcome of assessments is connected to employment consequences; I don't believe these challenges can be overcome without the costs outweighing the benefits. The costs include, for example, (a) the increased risk of instructional or curricular decisions being made to benefit employment rather than the students, (b) the increased risk that assessment data is corrupted as educators seek to demonstrate acceptable scores, and (c) the extreme dedication of time needed to develop assessments and structures that are defensible in

employment arbitration and also useful for all grade levels and content areas. I believe it would be much more productive to focus PE & PG systems on educators' analysis of data, encouraging educators to discover and collaboratively address challenges rather than to comply and potentially mask or avoid assessing areas of weakness. At the conclusion of this chapter I propose relevant avenues for practitioners, policymakers, and researchers.

Merit pay. The use of evaluation results to determine merit pay is not explicitly required by Maine statute and regulation, but is encouraged by language stating that "superintendents shall use effectiveness ratings of educators to determine strategic human capital decision making, including … compensation" (An Act to Ensure Effective Teaching and School Leadership, 2012). Merit pay is not in widespread use; Fairman and Mette (2017) found in a review of Department of Education data that in only 3% of the 146 responding districts did teachers earn financial compensation for ratings. In my study, from those without direct experience with merit pay, when participants raised the topic it was with skepticism or in the negative. In this study several had direct experience: one site was currently implementing merit pay, and a participant at another site had experience in a different district as a teacher receiving merit pay. That participant, now an evaluator, viewed merit pay poorly: "It's a sham…I think it caused more harm than good." Those interviewed at the site currently implementing merit pay were eager for the practice to end. The teacher interviewed (who was currently receiving great financial benefit from merit pay) shared specifics:

> I will say that teachers have become very good at manipulating the system to get the score they want, to get the money they want. I'm curious to see what this will look like when the money leaves this year. I wish the money was not available… I think once the money has gone there'll be great potential…It's had a negative impact. I don't know if it's

> decreased our professional performance. Certainly our ethics have decreased… I hear people talk about the scores they earn in anticipation of the money they're going to get, not in reflection of the performance they exhibited and where they can go.

The evaluator interviewed at the same site shared similar perspective, concluding that the use of merit pay brings more "yuckiness."

My study yielded limited data on merit pay; only a small number of participants had direct experience with such incentives. Keeping in mind the limited data, I saw themes emerging that were resonant with a 1986 exploration of why few teacher merit pay plans survive. There were many other facets to Murnane and Cohen's work (1986) on the topic, but one excerpt particularly connects to the teacher leader's statement in my study that ethics had decreased at her site: "If evaluation is to contribute to the goal of helping teachers improve, it must be carried out by skilled and knowledgeable supervisors in an atmosphere that rewards honesty and cooperation. When teachers who conceal their failings receive higher pay than those who do not, the atmosphere for useful evaluation and advice is poisoned" (p. 16).

Attempting to Improve Teacher Effectiveness Through Locally-Driven Changes

The second major theme of changes underway at the sites in my study were locally driven decisions that either operated in parallel to, or expanded upon what was strictly required by Maine's PE & PG mandates. These included work to expand classroom observation for evaluative and feedback purposes beyond the traditional model, and to shift culture to reduce teacher isolation and increase professional dialogue about teaching and learning practices.

Beyond clinical supervision. The traditional teacher evaluation described and criticized in the literature (e.g.; Hazi & Rucinski, 2009; Platt et al., 2000) involves an annual classroom observation using the clinical supervision model (pre-conference, observation, post-conference).

While actual practice may have varied at many districts, Weisberg et al. (2009) found 64% of tenured teachers in their study were observed two or fewer times during their most recent evaluation cycle, for an average total of 75 minutes. Maine's PE & PG mandate encouraged—but vague language did not strictly require—greater frequency: "[O]bservations of professional practice and formative feedback must occur each year and throughout the school year for all educators" (Chapter 180, p. 9). Similar to what Fairman and Mette (2017) found in a review of Maine's PE & PG system plans, most of the sites in this study maintained the familiar clinical supervision observation but aimed to increase the frequency of observations for evaluative and feedback purposes. More than half of the sites aimed to do so through adding brief observations (typically 10–20 minutes) followed by feedback (e.g., Marshall's mini-observations or Marzano's walkthrough observations using the iObservation platform). These efforts can contribute to the aims of Glickman et al.'s *collegial* schools if the feedback portion of the classroom visits is emphasized to deliberately develop knowledge and skills. Darling-Hammond (2013, p. 53) supports this approach: "[R]esearch has found that the frequent, skilled use of standards-based observation with feedback to the teacher is significantly related to student achievement gains, as the process helps teachers improve their practice and effectiveness."

In the qualitative data sites' efforts to expand observations for evaluative and feedback purposes were described positively, but the survey data conflicted. Respondents gave low results across most sites for "classroom observations are frequent enough for accurate teacher performance evaluation" and "ongoing constructive feedback is provided to teachers." In the survey data evaluators rated both of those indicators higher than teachers. There are multiple possible explanations, for example I speculate that the gap between the qualitative and quantitative findings may be in part the result of the early stages of piloting or implementing new

systems. The interview participants were typically Steering Committee members who, along with evaluators, are more on the front lines of piloting the new systems than most survey participants. Therefore, those respondents may have more experience with the intended expansion of observations for evaluation and feedback. Mason and Tu (2015) yielded similar data, finding in qualitative interviews of Maine evaluators at seven districts that they were largely satisfied with their observation systems.

Alternatively, the conflicting data may be the result of interview participants and evaluators conveying what they aspire to achieve, with the quantitative data showing what teachers have actually experienced. As I illustrated in the literature review, building on Weisberg et al.'s data, observations of even 1% of a Maine teacher's work with students would require far more classroom visits than any of the sites intended to achieve or actually achieved. The quantitative indicators "classroom observations are frequent enough for accurate teacher performance evaluation" and "ongoing constructive feedback is provided to teachers" may be difficult to satisfy without a massive expansion of classroom visits.

Whatever the explanation for the gap, the intended implementation of additional observations was more ambitious than evaluators were typically achieving. These planned expansions of classroom observations at the sites in my study were almost always without the addition of more supervisors or evaluators (one site added a district supervisor, another added a temporary grant-funded coach). Predictable time barriers emerged. One evaluator, for example, shared with regret that with the sheer number of teachers in his school he realistically hoped to achieve three or four brief observations per teacher rather than the intended ten. This is less frequent than recommended in the literature; Marshall (2013) seeks ten mini-observations and Marzano & Toth (2013) assert that sampling error is high with four or fewer observations. The

two teachers interviewed at the same site independently raised sympathy for the struggle. One said, "It's no different than most areas of teaching and education. It's just not enough time." The other, at a different school in the same district, sympathetically described the overall duties of his building's administrators and shared "administrators are nervous and worried in how the heck they're going to pull this off…they're just trying to stay up with day to day stuff. They're going out of their minds…the system is mind boggling." At a different site a supervisor shared, "[I]t feels like more is being asked of the people in those [administrative] positions. A lot more."

Observations less frequent than intended was potentially exacerbated by the state mandate that all educators be evaluated in the 2016–2017 year when I collected data; districts may move some personnel to multi-year cycles in future years. In total, sites had a modest expansion of classroom observations by evaluators underway, with potential then to yield a modest expansion of formative feedback and data collection for summative evaluation. But reaching the intended frequency of observations was often not yet occurring due to time demands on evaluators (including the ratio of teachers to evaluators and time required for documentation of observations). Further, sustainability was repeatedly raised as a concern even in the pilot phase where several sites were phasing in teachers or professional standards.

These time demands pose a great challenge to improved teacher effectiveness. Assigning aspirations—rather than resources—toward increased feedback and data collection will likely result long-term in supervision and evaluation systems with the same flaws criticized across the nation for decades. Without the dedication of tangible and sustainable resources toward supervision and evaluation, it is difficult to see how efforts to improve teacher effectiveness are to overcome typical inadequacies described in Chapter 2 (e.g., inter-rater reliability, inaccuracy, inconsistency, limited helpful feedback). In the next section I discuss attempts to improve

teacher effectiveness through cultural efforts less directly dependent on evaluator time, and then discuss practitioners' perceptions of their overall focus and the impact PE & PG implementation thus far has had on teacher effectiveness.

Cultural work to open doors. In addition to the planned additions of observations for evaluation and feedback, some sites were undertaking work to engage peer observers or change culture to achieve greater transparency and deliberate sharing of practices. One site, for example, deliberately trained peer observers in an effort to provide more feedback to teachers and remove some pressure from evaluators. In their vision for *collegial* schools, Glickman et al. (2013) discuss the need for growth-focused collaborative adult interactions improving school-wide teaching and learning, and quality instructional supervision that minimizes hierarchy. At half of the sites in my study practitioners were attempting to do those things through cultural work intended to increase trust and collaboration, encouraging teachers to open doors and engage with colleagues in discussion and examination of teaching and learning. These are positive steps toward *collegial* schools. Further, at more than half of the sites in my study districts had begun to implement the peer review component of PE & PG systems that is part of Maine's requirements (but for which districts have a great deal of local latitude). This is less than found by MEPRI; Fairman and Mette (2017) found in their review of state data that 82% of districts were utilizing peer observation in their PE & PG system plans. The peer review component also has potential to fit with these cultural efforts and greatly contribute to *collegial* schools, when implemented in a growth-focused rather than compliance focused way.

As presented in Chapter Four, participants at these sites positively viewed these cultural efforts; a teacher shared "It has given…me a lot of opportunity to learn from a lot of teachers that I would probably have never had direct contact with." Transition difficulties existed in building

these cultures and implementing peer review, but most of these appeared to be barriers that could be eased over time. Some teachers were hesitant to "put themselves out there" to share their strengths, and others needed to develop more trust in order to welcome colleagues into their classroom. One of the consistent barriers toward collaborative interaction that emerged in the study was shared by an evaluator: "The hard part is that teachers don't want to leave their classroom. They want to teach their kids." An administrative push to create job-embedded (Zepeda, 2006) collaborative interaction may be needed to overcome these transitional challenges.

These efforts that engage teachers and other staff in collaborative work focused on teaching and learning have great potential toward achieving Glickman et al.'s *collegial* schools. There is an opportunity to add formative feedback through these efforts, if done in a way well-planned to address the tendency toward leniency in peer observation (Marshall, 2013; Platt et al., 2000). Further, engaging colleagues in such conversations can help mitigate role tensions noted in the literature (e.g., Oliva & Pawlas, 2004). As most of the sites in my study do not employ non-evaluative supervisors administrators are typically playing both the evaluator and supervisor roles.

Attempting to Focus on Goals and Growth—Without Tangible Support

At the conclusion of each qualitative interview in this study I asked participants whether performance evaluation or professional growth had received more focus in their district. Across all sites growth was the dominant response, even with a state mandate textually more focused on evaluation. These qualitative data were backed up by quantitative results, with strong survey responses for "increased teacher professional goal setting" and "evaluators are supportive of teacher professional growth."

The data presented in Chapter Four, however, indicated more challenges than tangible increases in support for the growth aspirations, particularly at rural sites. Complications in teacher supervision were anticipated at rural sites due to role tensions where the only supervisor available may be the administrator (Oliva & Pawlas, 2004), and due to the relatively small number of staff who could possibly match the situational supervision needs described in Chapter 2 (e.g., instructional goals, the strengths and needs of the teacher, the career stage of the teacher, and the organizational goals). At three of the four rural sites in the study participants also shared great challenges with overall funding, staff recruitment, staff turnover (especially rapid administrator turnover), and geographic distance limiting professional development opportunities. This geographic challenge has been previously found by Eady and Zepeda (2007).

Just two sites—both non-rural—had palpable resource gains associated with the PE & PG system shift; another non-rural site had temporary additional funding and a temporary coach role that would soon disappear with the sunset of a Teacher Incentive Fund grant. Therefore most sites overall (and all rural sites) were attempting to yield increased effectiveness through existing resources. Below I share and discuss the results of those attempts.

Attempts to improve teacher effectiveness not yet successful. Those efforts to increase teacher effectiveness—largely through existing resources—were not yet resulting in an overall perceived increase in effectiveness. As described in Chapter Four, participants as a whole did not give greater ratings to current teacher effectiveness than under the previous system. The early stage of piloting or implementation and overall system complexity may explain why teacher effectiveness has not yet shifted. Some other key reasons teacher effectiveness had not yet changed are likely the barriers discussed earlier (e.g., time, resources, staff turnover).

Changing PE & PG systems was described by participants as complex, time-consuming work. Maine's 2012 law and subsequent statute and regulation changes affected several parts of chaptered law and added 17 pages of regulation with numerous details. Districts also needed to navigate local policy and local collective bargaining agreements in their PE & PG system work, and undertake a governance process that yielded the agreement of diverse stakeholder groups. The expenditure of time and energy to get to the piloting point was not trivial, and at the point of data collection system development and ratification may have consumed available capacity with actual instructional improvement yet to come.

I noticed during interviews that it was common for participants to need to refer to their PE & PG documents to recall process steps, and also noticed frequent self-corrections about system features. This unfamiliarity with a new, complex system may have played into the instructional feedback evaluators were willing to share with teachers early on; several authors in the literature have noted that evaluators limit potentially helpful or challenging feedback due to fear of litigation (DeSander, 2000; Glickman et al., 2013). The complexity challenge may persist past early implementation; one evaluator spoke candidly about this:

> There's so many loopholes…You make one little mistake, and it could be grievable, it could throw everything out. And because that creates such a complex system…There's a lot of room for error. So I think a lot of administrators, to play on the safe side, may not do what they should do. And I think I'll be tempted to do that too.

Though the general tendency to inflate evaluation ratings is well-documented in the literature (e.g., Bradshaw, 2002; Weisberg et al., 2009), I speculate that potential hesitation to share critical feedback may be at least temporarily exacerbated by Steering Committee members'

desire to generate teacher buy-in for the new PE & PG system or by fear of making an early procedural error (Sullivan & Zirkel, 1998) while rolling out a new, multifaceted system.

However, some subgroups in the study did perceive changes in teacher effectiveness. Evaluators, participants at the Danielson sites, and to a lesser extent participants at the Marshall sites gave stronger ratings to teacher effectiveness under the current system than the previous system. Conversely, the NBPTS sites gave the weakest ratings to the current system when compared to the previous system. These findings provide an important opportunity for follow-up study, which I will outline at the conclusion of this chapter.

Fundamentally, though, I return to time, which emerged as a major finding in my study and as a concern in the MEPRI studies. It has been well-recognized in previous studies that time is a limiting factor for teacher supervision and evaluation (e.g., Darling-Hammond, 2013; Goldstein, 2007; Marshall, 2013). Platt et al. (2000, p. 24) summarized the concern well: "Generally, administrators in American schools supervise and evaluate too many people annually to be able to do a creditable job." The changes underway at the sites in my study largely require more of the same educators. More is required of the teachers to engage with more detailed standards and rubrics, to provide greater evidence of their performance, to spend time on Student Learning Objectives, and to engage in collaborative discussion of teaching and learning. Where non-evaluative supervisors exist, those same changes exist as they assist each of their constituents. More is especially required of the administrators who typically provide both supervision and evaluation: more frequent observations, more detailed rubrics, more detailed feedback, more inclusive governance processes, and Student Learning Objectives for all of their teachers. I close the discussion by repeating one evaluator's exasperation:

> We have a great superintendent, we have great principals, and we are still struggling through this mess of trying to figure out what this system looks like. How on earth is this going to help teachers across the state of Maine and when will it get changed because people will find out how much work this is? … I'm just worried with very good intentions the state is requiring a system that is not sustainable, and so what's going to end up happening is a system that's just not carried out very well, that doesn't do what it was intended to do. Not because people don't care, but because there isn't time to do what's being asked well.

Concluding Thoughts and Implications

If one wished to improve teacher effectiveness at a school or school district, one would bring both evaluation and supervision to bear in a coordinated effort. Each would be built on the foundation of a comprehensive and agreed-upon set of professional standards. Through evaluation, one would assess individual and group performance, set goals, document and celebrate growth, and when needed—if the professional standards are not met—provide sufficient documentation for employment consequences (Sullivan & Zirkel, 1998). Through supervision, one would do what should be the much more frequent work of supporting teacher growth through such efforts as detailed instructional feedback, collegial dialogue about instruction, collaborative redesign of instructional plans, and modeling of excellent instructional practices. Ideally, that support would include a skilled supervisor that is not also the teachers' evaluator, to avoid role tensions and create conditions for effective support (Beach & Reinhartz, 1989; Oliva & Pawlas, 2004).

Beginning in 2012 Maine began passing a series of laws and regulations requiring systems of Performance Evaluation and Professional Growth, with the aim of increasing teacher

effectiveness. While more evaluation focused in the text, Maine sought to address both evaluation and supervision. Maine's PE & PG requirements gave sufficient latitude for school districts to design supervision well, using for example the principles laid out in Glickman, Gordon, and Ross-Gordon's well-regarded *collegial* model (2013). But, the requirements were laid upon school districts without substantial dedicated resources.

At the sites in my study changes were underway to pilot and implement Maine's PE & PG mandate. Those changes were not yet perceived to have had an overall impact on teacher effectiveness, which is not surprising at such an early stage of policy implementation. Maine school districts were largely exercising their latitude under the PE & PG mandate, choosing to undertake this work with a focus on formative teacher growth rather than summative evaluation and making some promising steps toward the *collegial* model. Early positive indicators include shared visions for teaching and learning, clear and useful standards and rubrics, and evaluator support for teacher growth. These factors and cultural work to open doors and increase collaboration contributed toward improved teacher effectiveness.

School districts still have much work to do, especially to increase the frequency of classroom observation and feedback for ongoing growth and for evaluative accuracy. At this stage of implementation participants in this study shared more barriers to improving teacher effectiveness than contributing factors. High on the list of barriers were evaluator time, teacher time, and the availability of non-evaluative staff to help teachers grow. Even in the pilot stage with a reduced number of participants, the intended frequency of classroom visits for instructional feedback and evaluation accuracy was not being reached. At rural sites, additional barriers existed in high turnover rates, attracting and retaining new staff, and geographic or financial barriers to accessing professional development.

More than 300 practitioners participated in my study: teachers, evaluators, and supervisors across Maine, across rural status, and across professional practice models. Through their efforts to design and implement PE & PG systems, Maine school districts appear to be reaching for the best of both worlds, attempting to implement an evaluation-focused mandate while largely focusing locally on professional growth. They are principally doing this through existing resources, seeking to increase observational frequency, instructional feedback, and support for growth with the same number of supervisors and evaluators they had before these changes. Unless one assumes that educators were sitting idle, needing a policy nudge to increase effort and become more effective, it is hard to see how unresourced PE & PG system changes will result in substantial sustained gains in teacher effectiveness. There are positive shifts underway at Maine school districts, but a fundamental challenge still exists: to realize the policy goal of increased teacher effectiveness, it appears the next step must be to supply more time for supervision and evaluation.

Implications for Practitioners

In order to make greater progress toward improved teacher effectiveness the findings indicate practitioners should: (1) continue reaping the benefits of the clear standards and rubrics, (2) continue building culture and sustained job-embedded practices (Zepeda, 2006; Disimone and Pak, 2017) to increase collegial discussion of teaching and learning, and (3) to the extent allowed by law minimize aspects of the PE & PG mandate that consume time with minimal impact on effectiveness.

First, practitioners have found the new standards and rubrics helpful and should use them in a growth-focused way, continuing to help educators deeply understand the standards and rubrics. For evaluators this means conceiving of supervision as an important part of their role,

and deliberately building skill in that area. For evaluators and supervisors this means dedicating time and attention to helping staff dig into the standards and rubrics, and frequently communicating how school efforts and initiatives align to the professional practices expressed in the model. This also means encouraging an environment in which all accurately describe rather than embellish the ratings of teacher practice. For teachers this also means dedicating effort to looking to the next rubric levels for guidance in goal-setting. Practitioners can continue to develop strategies or seek new resources to address the primarily time-based challenges slowing use of these comprehensive models for growth and evaluation. Practitioners should also recognize that at least transitional challenges to these comprehensive models are common (Derrington & Campbell, 2015) and thus exercise patience, share promising practices for growth and evaluation using the resources, and continue to dedicate resources to implementation.

Second, practitioners should recognize that cultural efforts and the shared vision of teaching and learning can quickly erode as other initiatives take the forefront and as new educators join the school district. Thus, ongoing reinvestment in building and maintaining *collegial* culture and a shared vision is warranted. Even in districts that employ non-evaluative supervisors, the circumstances may mean that different colleagues may be better suited to provide feedback and support relative to the situational needs (e.g., instructional goals, the strengths and needs of the teacher, the career stage of the teacher, and the organizational goals). For districts that do not employ non-evaluative supervisors, or for sites with frequent administrative turnover (e.g., most rural sites in my study), a shared cultural effort will be particularly important to yield gains amidst evaluator role tensions and the churn of new personnel.

Practitioners should learn and apply lessons from the literature in efforts to engage teachers and other staff in collaborative work and in giving instructional feedback, such as developing trainings and protocols in order to address the tendency toward leniency in peer observation (Marshall, 2013; Platt et al., 2000). Additionally, evaluators and supervisors can learn from the barriers identified in my study by seeking to foster feedback in a way that minimizes removing teachers from their students; job-embedded efforts (Zepeda, 2006) may yield good results. For teachers, while it is difficult to set aside the many immediate needs vying for attention, this means recognizing and treating such opportunities to share practices and feedback with each other as an essential and healthy part of a *collegial* school. As a caution, if merit pay is considered going forward those involved should deeply consider the possible unintended consequences on the culture of the school district.

Third, recognizing that increased teacher effectiveness is more likely to come through teacher growth rather than summative evaluation, practitioners should continue to find ways to streamline processes in whatever ways they can in order to maximize valuable feedback. From the findings in this study, this would include any efforts school districts can make to minimize unnecessary paperwork or documentation, maximizing instructional feedback. Practitioners may find benefit from shifting summative evaluation cycles to multi-year cycles in order to yield more time and attention for formative feedback. For evaluators, this means establishing methods by which to gradually collect evidence over multiple years, and to help others not lose sight of the standards and goals. For supervisors, this may mean coordinating coaching cycles with teachers' goal-setting and summative evaluation cycles. For teachers, this means designing ways to keep the standards, rubrics, and goals in mind even during non-summative years, so that continual progress is made, evidence gradually collected, and the pressure of the summative year

is mitigated. Further, within the confines of the law, practitioners should minimize energies spent toward elements locally viewed as not useful toward improving teacher effectiveness. For most districts in my study this would include minimizing energies spent toward the mandated Student Learning Objectives (SLOs), or altering the local approach to SLOs to yield greater *collegial* engagement with studying assessment data and identifying instructional implications.

Implications for Policymakers

In order to achieve the policy goal of improved teacher effectiveness the findings indicate policymakers: (1) should consider resources to increase the capacity of districts to provide ongoing formative feedback and growth opportunities to teachers, (2) should recognize the heavy dependence on administrators and seek ways to minimize time demands and reduce administrator turnover, and (3) should conduct a cost-benefit analysis of the current Student Learning Objectives mandate, informed by this and other emerging research.

First, recognizing that increased teacher effectiveness will come primarily from formative teacher growth rather than summative evaluation, policymakers should seek opportunities to contribute to this growth. One such development is underway with a proposal drafted to add more text to Chapter 180 regarding Peer Support and Mentoring, however the scarce time resource that emerged in my study illustrates the need to back the policy aspiration with resource. For policymakers this may mean efforts to add dedicated financial resource through which school districts can expand supervisory efforts for teachers (e.g., coaching, mentoring). Additionally or alternatively, policymakers can create state and regional opportunities to build upon commonalities of professional practice models and further support teacher growth, especially in ways that address the distance and financial barriers rural teachers face in accessing professional development. Policymakers should also recognize that some of educators' limited

time and attention resources must be dedicated to ongoing work to yield growth through PE & PG (Bradshaw, 2002; Darling-Hammond, 2013), and thus moderate or provide sufficient resources for new policy initiatives to facilitate maintaining or expanding attention to teacher growth. Bradshaw's (2002) description of how North Carolina's statewide evaluation effort faded once monitoring and resources decreased can serve as a cautionary tale to Maine policymakers.

Second, policymakers should be aware that even in the pilot stage evaluators are finding great time barriers in implementing PE & PG systems, which threaten the sustainability of these systems and presumably the overall effectiveness or retention of these educators. These time barriers were found in my study to impact achieving sufficient observations for evaluative accuracy and for formative feedback conversations, to impact promoting other school initiatives such as proficiency-based learning, and to impact the availability of administrators to help teachers with other issues. Thus in order to achieve the policy goal of increased teacher effectiveness additional personnel resources (evaluators or supervisors) will likely need to be brought into the picture (Derrington & Campbell, 2015). Further, policymakers must recognize that at sites with positive cultural work underway, the work was typically attributed to an administrator and thus seek whether there are any possible policy actions to reduce the factors leading to rapid administrative turnover (especially in rural school districts).

Third, the context under which Maine's 2012 Educator Effectiveness law and subsequent revisions were enacted has changed. The federal pressure to include student growth data in PE & PG systems dropped off when the *Every Student Succeeds Act* took the stage and *No Child Left Behind Flexibility Waivers* and *Race to the Top* passed into history. Thus, Maine policymakers have an opportunity to freely examine whether modifications to current statute and

regulation are warranted to further the policy aim of increased teacher effectiveness. In this examination policymakers should consider the findings from emerging research including my study and each MEPRI report (Mason & Porter, 2014; Mason & Tu, 2015; Mette & Fairman, 2016; Fairman & Mette; 2017). Reading these as a whole brings some satisfied perspectives from participants but also a list of persistent challenges in using student growth data (SLOs) in PE & PG systems (e.g., unintended consequences, gaming the system, validity questions, time expended). Perhaps most worrisome, in response to these tensions a number of participants in this study shared indicators resonant of those Argyris shared as warning signs inhibiting organizational learning: "error hiding, deception, and games…The moment individuals reach this state, they may also lose their ability to see the errors" (1977, p 4). Policymakers can determine whether or not the significant time currently invested in SLOs could be better used for other evaluation and growth activities such as providing more frequent instructional feedback.

Overall, policymakers have the opportunity to re-engage with the theory of action behind the PE & PG statutes and regulations. Leading up to the 2012 Educator Effectiveness law, federal pressure and widespread thinking about teacher evaluation is well-represented by Hallinger, Heck, and Murphy's depiction (2013, p. 8):

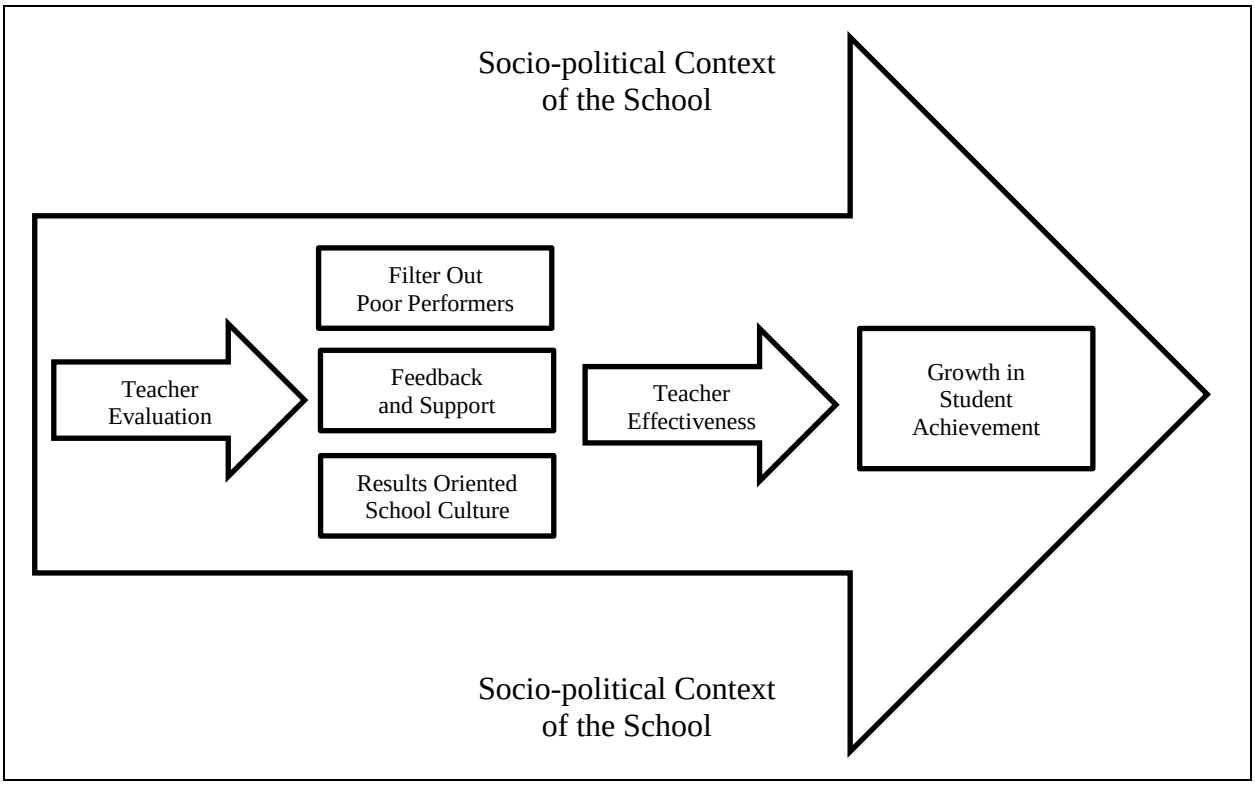

Figure 5.2. Theory of action underlying teacher evaluation and school improvement (Hallinger, Heck, & Murphy, 2013).

Hallinger et al. studied such "new generation" teacher evaluation systems, and found:

> This critical evaluation of the empirical literature yields two key conclusions. First, we conclude that the policy logic supporting this reform remains considerably stronger than the empirical evidence. Second, we suggest that alternative improvement strategies may yield more positive results and at a lower cost in terms of staff time and district funds (2013, p. 5).

Following the findings in my studies and others, I propose that policymakers consider a new theory of action for Maine's current context:

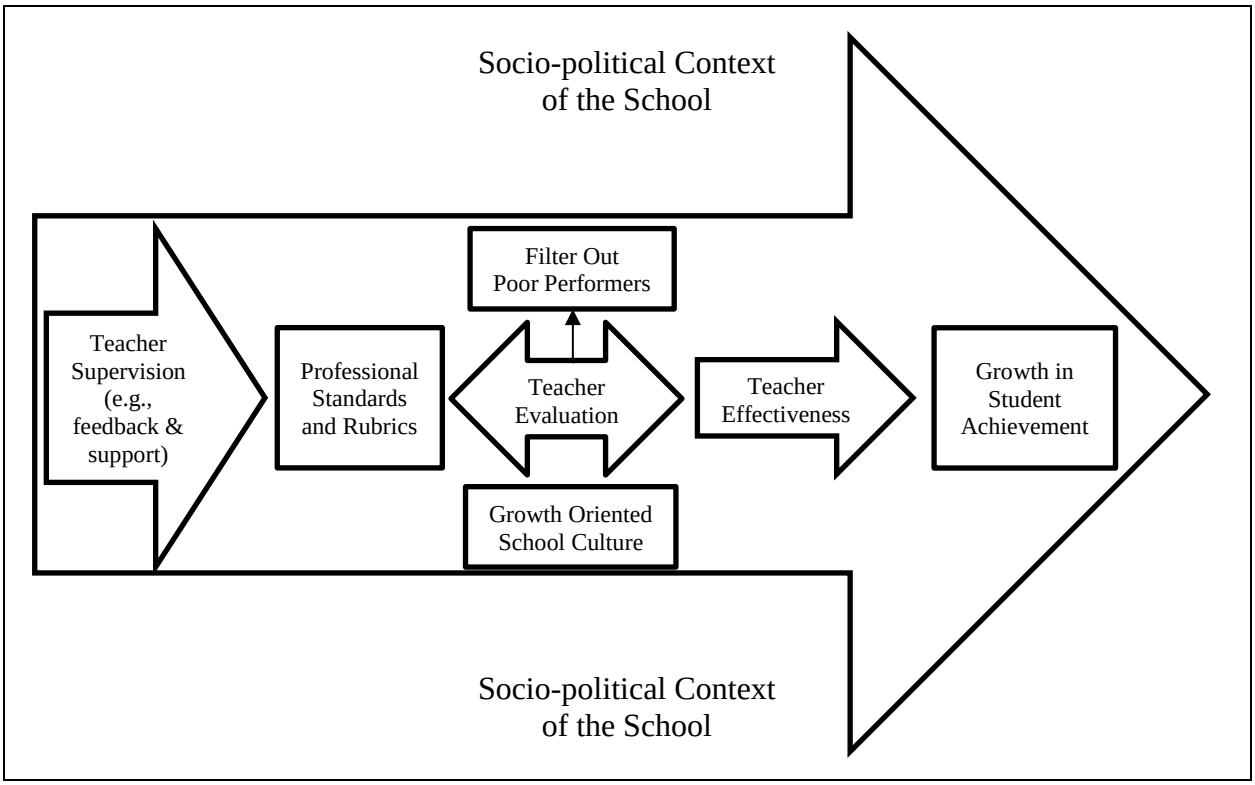

Figure 5.3. Proposed Maine theory of action underlying teacher supervision, evaluation, and school improvement. (Adapted from theory by Hallinger, Heck, & Murphy, 2013).

This proposed Maine theory of action builds upon the strengths found with the new professional practice standards and rubrics, but puts teacher supervision in the forefront of the model. Not only is supervision well-supported in the literature (e.g., Glickman et al., 2013), it simply fits the reality of many schools. Filtering out poor performers is a necessity in any field and remains in this proposed model. However, schools like those described in my study that struggle with staff turnover and finding certified and qualified applicants for open positions must focus on creating a growth-oriented culture and helping current and new teachers to thrive, so students may thrive.

Implications for Researchers

Researchers have a number of opportunities to build upon this research, including: (1) follow-up studies to explore how PE & PG system implementation changes after multiple years

of effort (such as a 3-year follow-up with a statewide survey and several deep qualitative case studies), (2) studies to compare and contrast the long-term impacts of choices where districts had some latitude (e.g., the professional practice model), and (3) studies focused more deeply on the controversial areas of student growth data and merit pay.

First, my study captured sites' PE & PG efforts at the pilot or initial implementation stage, revealing the changes underway at school districts and factors practitioners perceived as contributing to or barriers to increased teacher effectiveness. Follow-up studies after multiple years of PE & PG implementation could reveal how those efforts evolved, which efforts were sustained, if and how practitioners solved the challenges that emerged in the data, whether practitioners succeeded in their aspiration to focus on teacher growth rather than summative evaluation, and whether practitioners perceived a longer-term impact on increased teacher effectiveness.

Second, Maine provides fertile ground for researchers to compare and contrast PE & PG approaches. Researchers have an opportunity to compare and contrast ongoing implementation of four nationally-available models in the same state context, learning about the associations of these models and district demographics with diverse outcomes such as student achievement, teacher development, and educator retention. My study showed at this early stage that participants at the Danielson and Marshall sites gave stronger ratings to teacher effectiveness under the current system than the previous system, and that participants at the NBPTS sites gave the weakest ratings to the current system when compared to the previous system. Further study could determine if those results bear out in later implementation or across another sample. Researchers can also study the variety of approaches districts chose to implement the flexible peer review mandate and can contribute to knowledge about the conditions under which "open

door" cultures are or are not built and sustained, especially in districts with high teacher and administrator turnover.

Third, researchers can add to knowledge through focused study of merit pay and the use of student growth data in PE & PG systems. These factors were not the sole emphasis of my study but emerged as controversial and were generally viewed by practitioners as barriers to increased teacher effectiveness. With regards to the use of student growth data in PE & PG systems much data and opportunity for study will be available as practitioners are implementing these measures in a variety of ways, e.g., in every grade, in every content area, via standardized assessments, via locally-developed assessments, via individual measures, and via group measures. With regards to merit pay researchers have an opportunity to interview teachers, supervisors, and evaluators from a larger number of districts that have recently experienced merit pay (e.g., Teacher Incentive Fund districts). By doing so they can determine if the concerns raised by those in my study are typical or outliers.

Finally, Cuban (2011) described schools and classrooms as a "black box" in which little is known about how inputs are converted into outputs. Yielding increased educator effectiveness and increased student achievement – the policy goals behind PE & PG systems – no doubt brings into play the "complex mechanics and inter-relationships" to which Cuban refers. The literature on the topics of supervision and evaluation is weighted toward the theoretical rather than the empirical, toward frequently-tested subjects, and toward the perspectives of evaluators. Thus continued research is warranted on many fronts to continually increase and disseminate knowledge about what makes for effective learning, effective teaching, effective supervision, and effective evaluation.

REFERENCES

An Act to Ensure Effective Teaching and School Leadership. (2012). Maine Public Law, Chapter 635, LD 1858, 125th Maine State Legislature.

Argryis, C. (1977). Double loop learning in organizations. *Harvard Business Review, 55*(September-October), 115-124.

Baker, B., Oluwole, J., & Green, P. (2013). The legal consequences of mandating high stakes decisions based on low quality information: Teacher evaluation in the race-to-the-top era. *Education Policy Analysis Archives, 21*(5), 1–72.

Beckham, J. (1997). Ten judicial "commandments" for legally sound teacher evaluation. *West's Education Law Reporter, 117*, 435–439.

Bradshaw, L. K. (2002). Local district implementation of state mandated teacher evaluation policies and procedures: The North Carolina case. *Journal of Personnel Evaluation in Education, 16*(2), 113–127.

Brandon, J., Hollweck, T., Donlevy, J. K., & Whalen, C. (2018). Teacher supervision and evaluation challenges: Canadian perspectives on overall instructional leadership. *Teachers and Teaching, 24*(3), 1–18.

Brill, S. (2009, August). The rubber room: The battle over New York City's worst teachers. *The New Yorker*. Retrieved from http://www.newyorker.com

Chance, E. W. (1993). The Trojan Horse of educational reform: A look at one state's experience and the perceptions of selected school administrators. *Rural Educator*, *15*(1), 23–26.

Coladarci, T. (2003). *Gallup goes to school: The importance of confidence intervals for evaluating "adequate yearly progress" in small schools*. (Report for the Rural School and Community Trust). Retrieved from http://www.ruraledu.org/issues/nclb/coladarci.pdf

Creswell, J. W. (2008). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research.* Upper Saddle River, NJ.: Merrill/Prentice Hall.

Cuban, L. (2011). *Inside the black box of the classroom.* Retrieved from https://larrycuban.wordpress.com

Danielson, C. (2013). *The framework for teaching: Evaluation instrument.* Princeton, NJ: The Danielson Group.

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

Darling-Hammond, L. (2017). Teacher education around the world: What can we learn from international practice? *European Journal of Teacher Education, 40*(3), 291–309.

Dee, T. S., & Keys, B. J. (2004). Does merit pay reward good teachers? Evidence from a randomized experiment. *Journal of Policy Analysis and Management, 23*(3), 471–488.

Derrington, M. L., & Campbell, J. W. (2015). Implementing new teacher evaluation systems: Principals' concerns and supervisor support. *Journal of Educational Change*, *16*(3), 305–326.

DeSander, M. K. (2000). Teacher evaluation and merit pay: Legal considerations, practical concerns. *Journal of Personnel Evaluation in Education, 14*(4), 307–317.

Desimone, L. & Pak, K. (2017). Instructional coaching as high-quality professional development. *Theory Into Practice, 56*(1), 3–12.

Doore, B., Jorgensen, B., Doore, S., & Mason, C.A. (2013). *Teacher evaluation and professional growth systems: A national review of models, approaches, and challenges.* Retrieved from Maine Education Policy Research Institute website: http://mepri.maine.edu/

Donaldson, G. A. (2014). *From schoolhouse to schooling system: Maine public education in the 20th century.* Orono, Maine: Custom Museum Publishing LLC.

Donaldson, G. A. (2017, March 9). Improving educator effectiveness – useful feedback key to Maine's PEPG initiative. *Maine Schools in Focus.* Retrieved from https://umaine.edu/edhd/outreach/maine-schools-in-focus/

DuFour, R., & Eaker, R. (2009). *Professional learning communities at work: Best practices for enhancing student achievement*. Virginia: Solution Tree Press.

Fairman, J., & Mette, I. (2017). *Working toward implementation of performance evaluation and professional growth (PE/PG) systems in Maine school districts.* Retrieved from Maine Education Policy Research Institute website: http://mepri.maine.edu/

Glickman, C. D., Gordon, S. P., & Ross-Gordon, J. M. (2013). *SuperVision and instructional leadership: a developmental approach* (9th ed.). Boston: Allyn & Bacon.

Goldring, E., Grissom, J. A., Rubin, M., Neumerski, C. M., Cannata, M., Drake, T., & Schuermann, P. (2015). Make room value added: Principals' human capital decisions and the emergence of teacher observation data. *Educational Researcher*, *44*(2), 96–104.

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.

Hallinger, P., Heck, R. H., & Murphy, J. (2014). Teacher evaluation and school improvement: An analysis of the evidence. *Educational Assessment, Evaluation and Accountability, 26*(1), 5–28.

Hanushek, E. A., & Rivkin, S. G. (2006). Teacher quality. In E. Hanushek & F. Welch (Eds.), *Handbook of the Economics of Education, Volume 2* (1051-1078). St. Louis: Elsevier.

Hazi, H. M., & Rucinski, D. A. (2009). Teacher evaluation as a policy target for improved student learning: A fifty-state review of statute and regulatory action since NCLB. *Education Policy Analysis Archives, 17*(5).

Helge, D. I., & Marrs, L. W. (1981, April). *Recruitment and retention in rural America.* Paper presented at the National Conference on Special Education in Rural Areas, Murray, KY.

Holtzapple, E. (2003). Criterion-related validity evidence for a standards-based teacher evaluation system. *Journal of Personnel Evaluation in Education, 17*(3), 207–219.

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

Jimerson, L. (2005). Placism in NCLB—How rural children are left behind. *Equity & Excellence in Education, 38(3),* 211–219. doi: 10.1080/10665680591002588

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

Koppich, J. (2005). Addressing teacher quality through induction, professional compensation, and evaluation: The effects on labor-management relations. *Educational Policy, 19*(1), 90–111.

Koski, W. S. (2012). Teacher collective bargaining, teacher quality, and the teacher quality gap: Toward a policy analytic framework. *Harvard Law & Policy Review, Vol. 6,* 67–90.

Kowalski, T. J., & Dolph, D. A. (2015). Principal dispositions regarding the Ohio teacher evaluation system. *AASA Journal of Scholarship and Practice*, *11*(4), 4-20.

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

Lemke, J. C. (1994). Teacher induction in rural and small school districts. *Rural partnerships: Working together.* Texas: American Council on Rural Special Education.

Lomonte, C. (2015). *Intent to pilot survey results indicate SAUs are moving forward with PEPG systems*. Retrieved from Maine Department of Education website: https://mainedoenews.net/

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

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

Mason, C. A., & Tu, S. (2015). *Teacher professional evaluation and professional growth systems in Maine: 2015 report.* Retrieved from Maine Education Policy Research Institute website: http://mepri.maine.edu/

McKay, S. and E. Silva (2015). *Improving observer training: The trends and the challenges.* Princeton NJ: Carnegie Foundation for the Improvement of Teaching.

McMillan, J. H., & Schumacher, S. (2010). *Research in education: Evidence-based inquiry.* Boston, MA: Prentice Hall.

Mead, S., Rotherham, A., & Brown, R. (2012). The hangover: Thinking about the unintended consequences of the nation's teacher evaluation binge. *Teacher Quality 2.0. Special Report 2. American Enterprise Institute for Public Policy Research.*

Medley, D. M., & Coker, H. (1987). The accuracy of principals' judgments of teacher performance. *The Journal of Educational Research, 80*(4), 242–247.

Merriam-Webster. (2016). Dictionary. Retrieved from Merriam-Webster website: https://www.merriam-webster.com/

Mette, I., & Fairman, J. (2016). *Piloting PE/PG systems in Maine school districts: lessons learned*. Retrieved from Maine Education Policy Research Institute website: http://mepri.maine.edu/

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

Murnane, R. J., & Cohen, D. K. (1986). Merit pay and the evaluation problem: Why most merit pay plans fail and a few survive. *Harvard Educational Review, 56*(1), 1–18.

National Assessment of Educational Progress. (2009). NAEP 2008: Trends in academic progress. *National Center for Education Statistics, Vol. 2009-479,* 56.

National Center for Education Statistics. (2016). *Identification of rural locales.* Retrieved from https://nces.ed.gov/ccd/rural_locales.asp

Performance evaluation and professional growth systems. (2015). Maine Department of Education Chapter 180, under Title 20-A MRSA §13706.

Platt, A. D., Tripp, C. E., Ogden, W. R., & Fraser, R. G. (2000). *The skillful leader: Confronting mediocre teaching.* Acton, Massachusetts: Ready About Press.

Posey, L. (2016). *Summary of the Every Student Succeeds Act.* Retrieved from National Conference of State Legislatures website: http://www.ncsl.org/documents/educ/ESSA_summary_NCSL.pdf

Ramirez, A., Clouse, W., & Davis, K. W. (2014). Teacher evaluation in Colorado: How policy frustrates practice. *Management in Education, 28*(2), 44–51.

Reeves, C. (2003). *Implementing the No Child Left Behind Act: Implications for rural schools and districts.* Educational Policy Publications.

Rigby, J. G. (2015). Principals' sensemaking and enactment of teacher evaluation. *Journal of Educational Administration, 53*(3), 374–392.

Ritter, G. W., & Barnett, J. H. (2016). Learning on the job: Teacher evaluation can foster real growth. *Phi Delta Kappan, 97*(7), 48-52.

Ronfeldt, M., Loeb, S., & Wyckoff, J. (2013). How teacher turnover harms student achievement. *American Educational Research Journal, 50*(1), 4–36.

Saldaña, J. (2016). *The coding manual for qualitative researchers.* Los Angeles: Sage.

Schwartzbeck, T. D., Prince, C. D., Redfield, D., Morris, H., & Hammer, P. (2003). How are rural districts meeting the teacher quality requirements of No Child Left Behind? Charleston, VA: Appalachia Educational Laboratory.

Simmons, B. J. (2005). Recruiting teachers for rural schools. *Principal Leadership (High School Ed.), 5*(5), 48–52.

Spillane, J. P., Reiser, B. J., & Reimer, T. (2002). Policy implementation and cognition: Reframing and refocusing implementation research. *Review of Educational Research, 72*(3), 387–431.

Strong, M., Gargani, J., & Hacifazlioğlu, Ö. (2011). Do we know a successful teacher when we see one? Experiments in the identification of effective teachers. *Journal of Teacher Education, 62*(4), 367–382.

Stronge, J. H. (1997). Improving schools through teacher evaluation. In J. H. Stronge (Ed.), *Evaluating teaching: A guide to current thinking and best practice* (1–23). Thousand Oaks: Corwin.

Stronge, J. H., & Ostrander, L. P. (1997). Client surveys in teacher evaluation. In J. H. Stronge (Ed.), *Evaluating teaching: A guide to current thinking and best practice* (129–161). Thousand Oaks: Corwin.

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.

Sullivan, K. A., & Zirkel, P. A. (1998). The law of teacher evaluation: Case law update. *Journal of Personnel Evaluation in Education, 11*(4), 367–380.

Taylor, R. (1990). Interpretation of the correlation coefficient: a basic review. *Journal of Diagnostic Medical Sonography, 6*(1), 35–39.

Urban, W. J., & Wagoner, J. L. (2009). *American education: A history.* New York: Routledge.

Webb, P. T., & Gulson, K.N. (2013). Policy intensions and the folds of the self. *Educational Theory, 63*(1), 51-67.

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect.* Brooklyn, NY: The New Teacher Project.

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

Zepeda, S. J. (2012). *Instructional supervision: Applying tools and concepts* (3rd ed.). New York: Routledge.

APPENDIX A

QUALITATIVE SEMI-STRUCTURED INTERVIEW

Preliminary

1. Description of the study
2. Risks and Benefits of Participation
3. Informed consent
4. Permission to record

Pre-Interview Demographic Data (each participant)

1. Which of these best describes your primary responsibility in school?
    a. Teacher (providing direct instruction to students)
    b. Evaluator of teachers (e.g., Principal, Director)
    c. Supervisor of teachers (typically without evaluation responsibilities, e.g., instructional coach, literacy coach)
2. With what grade span(s) do you primarily work? (check all that apply)
    a. Primary Grades
    b. Upper Elementary Grades
    c. Middle School Grades
    d. High School Grades
3. Approximately how many students attend the school where you do most of your work? ___
4. How many full-time administrators serve the school where you do most of your work? ___
5. How would you describe your school district's current implementation of a Performance Evaluation and Professional Growth System for teachers?
    a. My district has not yet begun to develop a PE & PG System.
    b. My district has begun to develop a PE & PG System, but has not yet implemented most of it.
    c. My district has mostly developed a PE & PG System, and has begun to implement or pilot it.
    d. My district has mostly developed and implemented a PE & PG System.
    e. My district has fully developed and implemented a PE & PG System.
6. Which of these professional practice models is your school district using for teacher Performance Evaluation and Professional Growth?
    a. Danielson (a.k.a. *The Framework for Teaching)*
    b. InTASC
    c. Marshall, or the MSAD 49 model based on Marshall's work
    d. Marzano (a.k.a. *The Art & Science of Teaching Framework)*
    e. National Board for Professional Teaching Standards / Maine Schools for Excellence
    f. A locally-developed model (specify):_____
    g. Other (specify): _____

For the following questions, the initial question will be asked of all participants (e.g., #1). The interviewer will ensure that participants directly respond to the next level of questions (e.g., #1a, #1b), either through their initial response or as follow-up questions. The interviewer will ensure that participants either directly or indirectly address the final layer of questions (e.g., #1ai, #1bii).

1. Describe the major changes your school district has made to its teacher Performance Evaluation and Professional Growth System in the past several years.
    a. In what ways has your school's approach to teacher performance evaluation changed in the past several years?
        i. How has the focus on PE changed in the past several years?
        ii. How have the resources dedicated to PE changed?
    b. In what ways has your school's approach to teacher professional growth changed in the past several years?
        i. How has the focus on PG changed in the past several years?
        ii. How have the resources dedicated to PG changed?
2. Describe the impact teacher performance evaluation has had on teacher effectiveness at your school in the past several years.
    a. Describe recent local successes in improving teacher effectiveness via PE.
    b. Describe recent local challenges in improving teacher effectiveness via PE.
    c. Describe how you feel about the potential for improving teacher effectiveness at your school through PE.
3. Describe the impact teacher professional growth has had on teacher effectiveness at your school in the past several years.
    a. Describe recent local successes in improving teacher effectiveness via PG.
    b. Describe recent local challenges in improving teacher effectiveness via PG.
    c. Describe how you feel about the potential for improving teacher effectiveness at your school through PG.
4. Describe which has received more focus in your local district over the last several years, improving professional practice through professional growth or through performance evaluation.
    a. What experiences lead you to that conclusion?

APPENDIX B

QUANTITATIVE SURVEY INSTRUMENT

Teacher Performance Evaluation and Professional Growth in the Era of "Educator Effectiveness" in Maine

You are invited to participate in a research project being conducted by Jon Doty, a graduate student in the Educational Leadership doctoral program at the University of Maine (faculty sponsor: Dr. Ian Mette). The purpose of the research is to examine perspectives of local practitioners regarding early implementation of Maine's mandated systems of teacher performance evaluation and teacher professional growth (PE & PG systems).

What Will You Be Asked to Do?

If you decide to participate, you will be asked to complete an anonymous online survey, approximately 15 minutes in length. The survey questions will include demographics and questions about your local implementation of performance evaluation and professional growth systems.

Risks

Except for your time and inconvenience, there are few risks to you from participating in this study. There is the possibility that you may become uncomfortable answering the questions; you may skip any questions that you prefer not to answer. You may discontinue participation at any time.

Benefits

Potential benefits of this research include informing the field about performance evaluation and professional growth practices that are critical to teachers, leaders, and students. You may

additionally benefit from reflection on your own local efforts in teacher evaluation and teacher growth.

Confidentiality

The survey will be given through an online platform that uses industry-standard encryption to protect information. The survey will not ask you for your name, school name, or school district name; any open-ended responses will be carefully screened to be sure not to publish any identifiable information. Data will be password-protected or kept in the researcher's locked office, accessed by only him and his faculty advisor. The investigator may keep anonymous data indefinitely.

Voluntary

Participation is voluntary. If you choose to take part in this study, you may stop at any time. You may skip any questions you do not wish to answer. Return of the survey implies consent to participate.

Contact Information

If you have any questions about this study, please contact me at (*207-745-1715, doty@umit.maine.edu, or 156 Oak Street, Old Town ME 04468*). You may also reach the faculty advisor on this study at (*207-581-2733, Ian.Mette@maine.edu, or 334 Merrill Hall, University of Maine, Orono ME 04469*). If you have any questions about your rights as a research participant, please contact Gayle Jones, Assistant to the University of Maine's Protection of Human Subjects Review Board, at *207-581-1498 (or e-mail: gayle.jones@umit.maine.edu*).

Construct: Demographics

1. Which of these best describes your primary responsibility in school?
    a. Teacher (providing direct instruction to students)
    b. Evaluator of teachers (e.g., Principal, Director)
    c. Supervisor of teachers (typically without evaluation responsibilities, e.g., instructional coach, literacy coach)
2. With what grade span(s) do you primarily work? (check all that apply)
    a. Primary Grades
    b. Upper Elementary Grades
    c. Middle School Grades
    d. High School Grades
3. Approximately how many students attend the school where you do most of your work? ___
4. How many total full-time administrators serve the school where you do most of your work? For example enter 1.5 if you have a full-time principal and half-time assistant principal. ___

Construct: Performance Evaluation

The next several questions are about teacher **performance evaluation**, which is also called *summative evaluation*. Performance evaluation is typically a summary of performance over a specified period of time, usually for employment purposes (e.g., renewal of contract). On each, please choose if you strongly disagree, disagree, agree, or strongly agree with the statement.

5. Teacher performance evaluation in my school district is focused on a shared vision of teaching and learning. (SD/D/A/SA)
6. In my school district, the amount of teacher time needed for the teacher performance evaluation process is reasonable. (SD / D / A / SA)
7. In my school district, the amount of evaluator time needed for the teacher performance evaluation process is reasonable. (SD / D / A / SA)
8. In my school district, classroom observations are frequent enough for accurate teacher performance evaluation. (SD / D / A / SA)
9. In my school district, clear standards exist for teacher performance evaluation (SD / D / A / SA)
10. In my school district, clear rubrics or scales describe the standards for teacher performance evaluation (SD / D / A / SA)
11. In my school district, the standards for teacher performance evaluation are relevant to teachers' responsibilities (SD / D / A / SA)
12. In my school district, sufficient training is available for teachers to understand the process of performance evaluation. (SD / D / A / SA)
13. In my school district, sufficient training is available for evaluators to understand the process of teacher performance evaluation. (SD / D / A / SA)

14. In my school district, the process of combining measures into a summative effectiveness rating is clear. (SD / D / A / SA)
15. In my school district, the summative effectiveness rating accurately reflects teacher effectiveness. (SD / D / A / SA)
16. In my school district, the summative effectiveness rating distinguishes between levels of teaching effectiveness. (SD / D / A / SA)
17. In my school district, teachers were involved in developing the standards for teacher performance evaluation. (SD / D / A / SA)
18. In my school district, teachers were involved in developing the process for teacher performance evaluation. (SD / D / A / SA)
19. In my school district, teachers are involved in the ongoing governance of teacher performance evaluation (for example, “Steering Committee”). (SD / D / A / SA)

20. What are the most major changes your school district has made in teacher performance evaluation over the past several years? (rank order the top five choices)

    a. Changed professional practice standards
    b. Developed rubrics or scales for professional practice standards
    c. Increased training about professional practice standards
    d. Increased training about the evaluation process
    e. Increased frequency of full-class observations
    f. Increased frequency of "mini" or "walkthrough" observations
    g. Increased frequency of unannounced observations
    h. Increased frequency of pre-planned observations
    i. Increased frequency of summative evaluations
    j. Increased number of people with evaluation responsibilities
    k. Increased use of evaluation outcomes in collaboratively-set professional goals
    l. Increased use of evaluation outcomes in evaluator-set professional goals
    m. Increased use of favorable evaluation results for opportunities a person would typically desire
    n. Increased use of unfavorable evaluation results for consequences a person would typically not desire
    o. Increased use of student growth data in evaluation
    p. Other (specify):

Construct: Professional Growth

The next several questions are about teacher **professional growth**, a formative and ongoing process focused on teacher development. Efforts toward **professional growth** are often varied: formal and informal, guided and self-directed, individual and group. On each, please choose if you strongly disagree, disagree, agree, or strongly agree with the statement.

21. Teacher professional growth in my school district is focused on a shared vision of teaching and learning. (SD/D/A/SA)
22. In my school district, collaborative teacher professional growth is well-supported. (SD / D / A / SA)
23. In my school district, individual teacher professional growth is well-supported. (SD / D / A / SA)
24. In my school district, teacher professional growth is well-supported within the school schedule / calendar. (SD / D / A / SA)
25. In my school district, teacher professional growth is well-supported outside of the school schedule / calendar. (SD / D / A / SA)
26. In my school district, non-evaluative staff are available to help teachers grow (e.g., literacy coach, instructional coach). (SD / D / A / SA)
27. In my school district, ongoing constructive feedback is provided to teachers. (SD / D / A / SA)
28. In my school district, teachers are supported in reflective inquiry. (SD / D / A / SA)
29. In my school district, evaluators (e.g., principal, director) are supportive of teacher professional growth. (SD / D / A / SA)

30. What are the most major changes your school district has made in teacher professional growth over the past several years? (rank order the top five choices)

    a. Increased funding for workshops or conferences
    b. Increased funding for university coursework
    c. Increased teacher support via coaching
    d. Increased teacher peer support and/or peer collaboration
    e. Increased teacher professional goal setting
    f. Increased support for teacher professional goals
    g. Increased time for professional growth embedded in the school schedule (e.g., early releases, in-service days, within meeting schedule)
    h. Increased instructional feedback through short classroom visits (a.k.a. "mini-observations"
    i. Increased instructional feedback through full-class visits / observations
    j. Increased use of a professional development software system
    k. Increased training and/or collaboration on analysis of student data
    l. Increased collaboration on analysis of student data
    m. Increased emphasis on reflective inquiry
    n. Differentiated professional development
    o. Other (specify):

Construct: Changes in Perceived Teacher Effectiveness

31. Prior to Maine's PE & PG law (2012 and earlier), overall teacher effectiveness was strong in my school district. (SD / D / A / SA)
32. Overall teacher effectiveness is currently strong in my school district. (SD / D / A / SA)
33. Prior to Maine's PE & PG law (2012 and earlier), my school district had a system of teacher performance evaluation that improved overall teacher effectiveness. (SD/D/A/SA)
34. My school district currently has a system of teacher performance evaluation that improves overall teacher effectiveness. (SD/D/A/SA)
35. Prior to Maine's PE & PG law (2012 and earlier), my school district had a system of teacher professional growth that improved overall teacher effectiveness. (SD/D/A/SA)
36. My school district currently has a system of teacher professional growth that improves overall teacher effectiveness. (SD/D/A/SA)

Construct: Influencing Factors

37. Over the last several years, what factors in your school district have helped to improve overall teacher effectiveness through performance evaluation and/or professional growth?
38. Over the last several years, what factors in your school district have been barriers to improving overall teacher effectiveness through performance evaluation and/or professional growth?

APPENDIX C

LETTER TO PARTICIPANTS

This letter is adapted from a University of Maine Institutional Review Board sample, accessed May 30th, 2016.

Dear _____,

You are invited to participate in a research project being conducted by Jon Doty, a graduate student in the Educational Leadership doctoral program at the University of Maine (faculty sponsor: Dr. Ian Mette). The purpose of the research is to examine perspectives of local practitioners regarding early implementation of Maine's mandated systems of teacher performance evaluation and teacher professional growth (PE & PG systems).

What Will You Be Asked to Do?

(Version A, qualitative interview): If you decide to participate, you will be asked to take part in an individual interview, approximately 45-60 minutes in length. The interview will be audio-recorded and will include demographic questions as well as prompts such as:

(1) Describe the major changes your school district has made to its teacher Performance Evaluation and Professional Growth System in the past several years.

(2) In what ways has your school's approach to teacher professional growth changed in the past several years?

(Version B, quantitative survey): If you decide to participate, you will be asked to complete an anonymous online survey, approximately 15 minutes in length. The survey questions will include demographics and questions about your local implementation of performance evaluation and professional growth systems. Example questions include:

(1) What are the most major changes your school district has made in teacher professional growth over the last several years?

(2) In my school district, clear rubrics or scales describe the standards for teacher performance evaluation. (Strongly Disagree / Disagree / Agree / Strongly Agree)

Risks

Except for your time and inconvenience, there are few risks to you from participating in this study. There is the possibility that you may become uncomfortable answering the questions; you may skip any questions that make you uncomfortable. You may discontinue your participation at any time.

Benefits

Potential benefits of this research include informing the field about performance evaluation and professional growth practices that are critical to teachers, leaders, and students. You may additionally benefit from reflection on your own local efforts in teacher evaluation and teacher growth.

Confidentiality

(Version 1, qualitative interview): Following the interview the researcher may temporarily use the assistance of a skilled transcriptionist to transcribe the audio recordings, who will agree to treat the data confidentially. Your name, your school's name, or other identifying information will not be reported in any publications. Your name will not be on any of the data; a code name or number will be used to protect your identity. Data will be password-protected or kept in the researcher's locked office, accessed by only him and his faculty advisor. Any recordings or documents with names, and the key linking names to the data will be destroyed

after data analysis is complete in early 2017; the investigator may keep anonymous data indefinitely.

(Version 2, quantitative survey): The survey will be given through an online platform that uses industry-standard encryption to protect information. The survey will not ask you for your name, school name, or school district name; any open-ended responses will be carefully screened to be sure not to publish any identifiable information. Data will be password-protected or kept in the researcher's locked office, accessed by only him and his faculty advisor. Any recordings or documents with names, and the key linking names to the data will be destroyed after data analysis is complete in early 2017; the investigator may keep anonymous data indefinitely.

Voluntary

Participation is voluntary. If you choose to take part in this study, you may stop at any time. You may skip any questions you do not wish to answer. Return of the survey implies consent to participate.

Contact Information

If you have any questions about this study, please contact me at (*207-745-1715, doty@umit.maine.edu, or 156 Oak Street, Old Town ME 04468*). You may also reach the faculty advisor on this study at (*207-581-2733, Ian.Mette@maine.edu, or 334 Merrill Hall, University of Maine, Orono ME 04469*). If you have any questions about your rights as a research participant, please contact Gayle Jones, Assistant to the University of Maine's Protection of Human Subjects Review Board, at *207-581-1498 (or e-mail gayle.jones@umit.maine.edu).*

BIOGRAPHY OF THE AUTHOR

Jonathan E. Doty was born in Florence, Alabama and raised in Wescosville, Pennsylvania. He graduated from Emmaus High School (Emmaus, PA) in 1996, then pursued post-secondary education at the University of Maine (Orono, ME). At the University of Maine he earned a Bachelor's Degree in Elementary Education with a concentration in Natural Sciences (2000), a Master's Degree in Education with a concentration in Instructional Technologies (2004), and a Certificate of Advanced Study in Education (2006) focused on Gifted and Talented Education and Educational Leadership.

Professionally, Doty has served Old Town School Department (Old Town, ME) and Regional School Unit #34 (Alton, Bradley, and Old Town, ME) in a variety of roles: Technology Education Teacher; Middle School Math & Science Teacher; Coordinator of Gifted & Talented Services; and Director of Curriculum, Instruction, and Assessment. Doty has also served as Director for several youth summer camps in that time, and worked part-time in emergency medicine and Maine's whitewater rafting industry.

He is a candidate for the Doctor of Education degree in Educational Leadership from The University of Maine in May 2018.