AN INTERPRETATIVE PHENOMENOLOGICAL ANALYSIS EXPLORING TEACHER EXPERIENCES OF EDUCATOR EVALUATION IN MASSACHUSETTS

A thesis presented
by

Jill Taylor Story

In partial fulfillment of the requirements for the degree of
Doctor of Education

in the field of
Education

College of Professional Studies
Northeastern University
Boston, Massachusetts
November 8, 2017
Abstract

Public education reforms that attempt to more accurately determine the impact of individual teachers on student learning outcomes have resulted in recent changes to educator evaluation systems. In Massachusetts, teachers are evaluated according to their performance on the standards of effective teaching and their progress on student learning and professional practice goals. Using the Concerns-Based Adoption Model framework, this study explored teacher experiences with educator evaluation to understand how teachers have been impacted by externally imposed regulations related to their practice. The data, which consisted primarily of semi-structured interviews, was analyzed using an interpretative phenomenological analysis. The findings reveal that teachers appreciate having a structured system of evaluation; however, they are not aware how to utilize the results from standardized tests to appropriately revise curriculum and instruction to improve their own performance ratings. Additionally, educators’ experiences with evaluation depend largely on their relationship with their evaluator and the ways in which evaluation procedures have been implemented in districts and schools. A lack of consistency in implementation policies and varying evaluator expectations, therefore, indicate that educator performance ratings reported to the Massachusetts Department of Elementary and Secondary Education are unreliable.

Keywords: educator evaluation, effective teaching practices, teacher accountability, educator performance rating
Acknowledgments

When I first embarked on the goal of earning a doctorate, I could never have imagined the ways in which this journey would challenge me mentally and emotionally. The pressures from school, work, and family seemed overwhelming at times. Through it all, my loved ones demonstrated patience and understanding when I needed to finish readings for a class, post to a discussion board, or research and write. I am most appreciative of the support from my husband, Brian, and my children, Jacqueline, Brian, Jr., and Alexandra. I would also like to extend my gratitude to my parents, who fostered in me a love of learning and instilled in me the values of hard work and perseverance.

I would like to thank my peers at Northeastern; those who I met at residency and those who kept me moving forward when I felt like giving up. I am incredibly grateful for the positive guidance and support from my advisor, Dr. Karen Reiss Medwed, who graciously took me on as an advisee during a period of transition and uncertainty. Her encouragement and feedback were crucial to my success. I would also like to thank Dr. Chris Unger and Dr. Dorothy Flaherty for their time and feedback on my work.
# TABLE OF CONTENTS

ACKNOWLEDGEMENTS..................................................................................................................3

CHAPTER 1: INTRODUCTION ......................................................................................................7

  Statement of the Problem..............................................................................................................7
  Research Problem .........................................................................................................................8
  Justification for the Research Problem ........................................................................................12
  Deficiencies in the Evidence........................................................................................................13
  Relating the Discussion to Audiences ..........................................................................................13
  Significance of the Study ............................................................................................................14
  Positionality Statement ................................................................................................................16
  Purpose Statement and Research Question ..................................................................................19
  Theoretical Framework ................................................................................................................20

CHAPTER 2: LITERATURE REVIEW ............................................................................................22

  Historical Perspective on Teacher Evaluation Systems.............................................................23
  Student Growth Models ..............................................................................................................32
  Educational Impact of Value-Added Teacher Accountability ......................................................40
  Limitations ..................................................................................................................................50
  Summary .....................................................................................................................................51

CHAPTER 3: METHODOLOGY ...................................................................................................54

  Purpose Statement and Research Question ..................................................................................54
  Research Design ..........................................................................................................................54
  Research Tradition ......................................................................................................................56
  Participants ...................................................................................................................................59
Recruitment and Access ................................................. 60
Data Collection ................................................................... 61
Data Storage ....................................................................... 62
Data Analysis ....................................................................... 62
Ensuring Trustworthiness and Verification ................................ 64

CHAPTER 4: RESEARCH FINDINGS ....................................... 67
Context ................................................................................. 67
Participants ........................................................................... 68
Themes .................................................................................. 75
Conclusion ............................................................................. 101

CHAPTER 5: DISCUSSION OF THE RESEARCH FINDINGS .......... 103
Revisiting the Problem of Practice ........................................ 103
Review of Methodology ....................................................... 104
Discussion of Key Findings ................................................... 105
Discussion of Findings in Relation to the Theoretical Framework ........................................................................ 110
Discussion of Findings in Relation to the Literature Review ................................................................. 115
Implications for Practice ........................................................ 117
Limitations ............................................................................... 119
Significance of the Study ....................................................... 119
Conclusion ............................................................................. 120
References ............................................................................ 122
Appendix A – Teacher Rubric ................................................. 137
Chapter 1: Introduction

Statement of the Problem

Ask any veteran educator to identify the qualities that describe a skillful teacher and they are likely to respond with characteristics like “compassionate,” “dedicated,” and “caring.” In recent years, however, these attributes, which are essential for forging the trusting relationship between teacher and student that is conducive to learning, are virtually absent from educator evaluation requirements in K-12 public schools. This trend began when national and state policies shifted at the turn of the century to place greater accountability for student achievement on schools, and continued with the implementation of revised teacher evaluation systems that emphasize the contributions of individual teachers to student learning. Truly, policy changes have significantly impacted teachers, students, and school administrators in a myriad of ways.

The passage of the No Child Left Behind Act (NCLB) in 2001 and the Race to the Top (RTTT) initiative of 2009, which established national expectations for public schools that determined school quality and teacher effectiveness based significantly on student test score results, expanded federal involvement in the realm of public education. NCLB effectively compelled states to implement and report on high-stakes testing outcomes and to tighten up licensing requirements for educators to ensure that schools employed “highly qualified teachers.” Schools that failed to make Adequate Yearly Progress (AYP) under NCLB could face sanctions or restructuring (Liston, Whitcomb, & Borko, 2007).

While the alleged purpose of NCLB was to close the achievement gap by ensuring that students from all racial, cultural, and socioeconomic backgrounds had
access to a high-quality education, these objectives were not achieved under NCLB since the law measured student achievement according to performance on “standardized tests that all students [were] required to take, in the same way and at the same time” (Gay, 2007, p. 281). The standardization and high-stakes testing requirements under NCLB contradicted the aims of the legislation to provide educational access to diverse students with a range of abilities. Furthermore, the constraints imposed on teachers under NCLB hindered, rather than enhanced, the “professional efficacy, motivation, and imagination” (p. 280) of educators.

Not only were teachers negatively impacted by NCLB requirements, the tracking of student performance as a condition of NCLB was also problematic since AYP was reported using a status model, comparing students from a current year with those from the previous year. In 2006, the Department of Education allowed a select number of states to pilot the use of growth models, designed to measure the learning gains of a cohort of students over several years, as an alternative (Mathis, 2006). These status models would eventually be incorporated into the competitive grant program known as RTTT, which was developed under the American Recovery and Reinvestment Act of 2009. RTTT incentivized states to focus on school turnaround strategies for underperforming schools and improvements in educator evaluation in exchange for more than $4.35 billion in funding. In an effort to comply with RTTT, many states, including Massachusetts, adopted Common Core State Standards and revised teacher evaluation systems to incorporate student achievement data in the assessment of teachers (Amrein-Beardsley & Collins, 2012; Ballou & Springer, 2015).

**Research Problem**
In June 2011, the Massachusetts Department of Elementary and Secondary Education adopted a new regulatory framework for educator evaluation to comply with the conditions outlined for RTTT states. The new regulations led to the development of a Model System for Educator Evaluation designed to improve the effectiveness of the teacher evaluation process. Originally released in January 2012, the changes to educator evaluation were formulated with the intent of “strengthening professional practice and improving student learning” through “analytical conversation about teaching and leading” (Massachusetts Department of Elementary and Secondary Education, 2012). Developed with input from key stakeholders, the Model System for Educator Evaluation was comprised of the following eight components to support the implementation of the new regulations:

I. District-Level Planning and Implementation Guide
II. School-Level Planning and Implementation Guide
III. Guide to Rubrics and Model Rubrics for Superintendent, Administrator and Teacher
IV. Model Collective Bargaining Contract Language
V. Implementation Guide for Principal Evaluation
VI. Implementation Guide for Superintendent Evaluation
VII. Rating Educator Impact on Student Learning Using District-Determined Measures of Student Learning
VIII. Using Staff and Student Feedback in the Evaluation Process

The Department of Elementary and Secondary Education informed districts that they could “adopt or adapt the Model System or revise their own evaluation systems to meet
the educator evaluation regulation” (Massachusetts Department of Elementary and Secondary Education, 2012). The expectation was that district leaders would be in charge of specific evaluation requirements for their schools, and building leaders would implement evaluation procedures using common rubrics and language established by DESE.

Per the regulations, the current system of evaluation involves a five-step process, consisting of self-assessment, goal setting and plan development, plan implementation, mid-cycle review, and summative evaluation, and teachers are either placed on a Self-Directed Growth Plan, Directed Growth Plan, Improvement Plan, or Developing Educator Plan depending on their experience and performance. Each of these plans requires varying degrees of involvement from the evaluator, as well as considerably more work for the educator and evaluator than previous evaluation methods. Under the Model System, teachers are evaluated on the following standards:

- *Curriculum, Planning and Assessment*
- *Teaching All Students*
- *Family and Community Engagement*
- *Professional Culture*

Additionally, three categories of evidence are incorporated into the assessment of educator performance:

- **Multiple measures of student learning, growth, and achievement**, including classroom assessments, district-determined measures comparable across grade or subject district-wide, and state-wide growth measures where available, including
the MCAS Student Growth Percentile (SGP) and Massachusetts English Proficiency gain scores (MEPA);

- **Judgments based on observation and artifacts of professional practice,** including unannounced observations of practice of any duration; and

- **Additional evidence relevant to one or more Performance Standards,** including student feedback as a source of evidence when evaluating teachers and administrators, and staff feedback when evaluating administrators (Massachusetts Department of Elementary and Secondary Education, 2012).

Educator performance is rated as *Exemplary, Proficient, Needs Improvement,* or *Unsatisfactory* based on the Performance Standards listed above. Until the amending of the regulations in early 2017, teachers were also issued an Impact on Student Learning rating of high, moderate, or low “based on trends and patterns in learning gains on state and district-determined measures of student learning, growth, and achievement” (Massachusetts Department of Elementary and Secondary Education, 2012). Beginning in 2013, districts reported both the Educator Performance Rating and the Impact on Student Learning Rating to DESE. The Impact on Student Learning Rating was discontinued during the 2016-2017 school year, and new regulations specify that student learning should be embedded as an indicator within Standard II of the evaluation framework beginning in the 2017-2018 school year, although DESE has not provided instructions yet for how to do this (Massachusetts Department of Elementary and Secondary Education, 2017).

Indeed, the new evaluation model signifies a departure from the old model based primarily on supervisor observations, due in part to research that suggests that
supervisors’ classroom observations alone are not accurate predictors of teacher quality (Strong, Gargani, & Hacifazlioglu, 2011). In Massachusetts, a RTTT state, changes to evaluation procedures compel teachers to demonstrate their impact on student learning in ways that are quantifiable. The five-step process for evaluation places increasing requirements on evaluators and educators in terms of goal setting, evidence submission, and reporting of educator performance. As a result, teachers have had to adapt their professional practices in order to implement externally imposed requirements to prove they are at least minimally proficient.

**Justification for the Research Problem**

While at first glance, these changes may seem positive, exposing both effective and ineffective teachers, research reveals that evaluative tools that attempt to isolate teacher impact are not necessarily dependable measures of teacher effectiveness (Darling-Hammond, Amrein-Beardsley, Haertel, & Rothstein, 2012; Rothstein, 2010). Therefore, in the haste to revise educator evaluation to more accurately reflect student achievement, school districts that evaluate teachers based on student learning gains risk using faulty measures to address characteristics of effective teaching. Furthermore, teachers may be aware of validity concerns with evaluations based on student achievement metrics, but are evaluated according to these standards nonetheless (Amrein-Beardsley & Collins, 2012; Papay, 2011; Rothstein, 2009).

Much of the research to date focuses on how teachers are evaluated using statistical models for student achievement, or compares these more objective measures with other indicators of teacher effectiveness, such as personality or relatability (Jacob & Lefgren, 2008; Strong et al., 2011). Additionally, there are aspects of teacher
effectiveness, such as empathy and compassion for students, which are not easily quantifiable. The existing research often fails to account for alternative measures of teacher quality or consider how teachers have experienced changes to educator evaluation practices.

**Deficiencies in the Evidence**

Missing from the research are the experiences and perceptions of the teachers who have had to adapt their practices in order to demonstrate their impact in a quantifiable manner, or who define teacher quality in a way that is not easily measured. Therefore this study seeks to understand how middle school teachers of state-tested subject areas in Massachusetts have experienced changes to evaluation procedures that rate educator performance based on standards and student achievement and the effect of these methods on teaching attitudes. Through better understanding the impact of evaluation changes on the teaching profession, administrators will be able to refine their evaluation practices to make the process more authentic and gratifying for evaluators and educators, as well as identify strategies for supporting continued professional growth. Additionally, policymakers will better comprehend how top-down accountability measures affect teachers in the classroom.

**Relating the Discussion to Audiences**

This study will help administrators and policymakers better understand how laws and regulations affect the educators who are charged with implementing changes in the classroom. Too often, well-intentioned policies have unintended consequences that negate their broader purpose. Research into the impact of high-stakes evaluation policies
on key stakeholders is essential for shedding light on challenges not anticipated and informing future policy decisions.

The research will benefit classroom teachers by illuminating their experiences so that leaders will either modify existing policies, or at the very least, be capably equipped with information to appropriately support teachers with sufficient professional development and resources to promote self-efficacy. Teachers thrive in a supportive environment, and if existing educator evaluation models are here to stay, school leaders should understand how teachers view the procedures in order to assist them with navigating the requirements. Insight into the experiences of educators will enable administrators to anticipate potential difficulties with evaluation and better respond to teacher concerns with targeted strategies to alleviate any issues.

**Significance of the Study**

This qualitative research study will investigate how reliance on performance standards and ratings for the evaluation of educators impacts the teaching profession. Increased standardization and accountability measures for schools, which are the result of encroaching national and state policies, have educational consequences. Studies have shown that an overemphasis on testing, which is entwined with evaluation, may negatively impact the curriculum, teaching, and learning, especially for certain student populations, such as low-income, minority, learning disabled, or English Language Learners (Agee, 2004; Amrein-Berliner, 2003; Costigan, 2002; Darling-Hammond, 2007; Diamond & Spillane, 2004; Hollingworth, 2009; Horn, 2003; Jones, 2007; Polesel, Dulfer, & Turnbull, 2012; Scherrer, 2012). Furthermore, Lee (2011) found that often the teachers, as the individuals who are most impacted by educational policy decisions, are
left out of the conversations pertaining to school reform. Indeed, the use of performance ratings and reliance on student achievement data to evaluate educators undermines the larger purpose many teachers assign to education (Everson, Feinauer, & Sudweeks, 2013; Lee, 2011).

The centralization of education policy has had a significant impact at the national, state, and local level. At the national level, the NCLB mandate required states to establish annual targets for schools to make AYP, which often varied across states. Additionally, the content and rigor of tests used to determine student proficiency has been inconsistent, making it nearly impossible to compare state educational systems based on AYP data (Usher, 2012). When 48% of the nation’s schools did not make AYP in the 2010-2011 school year, U.S. Secretary of Education Arne Duncan issued NCLB waivers to provide states with an opportunity to devise alternative measures for accountability that, if approved by the federal government, could be used to raise standards instead of APY (Usher, 2012). This shift in national policy prompted school leaders to modify existing practices of teacher evaluation to ensure compliance with changing federal conditions (Ballou & Springer, 2015).

In many public school districts, state and federal accountability policies have led to an alignment of the curriculum with standards, a heightened focus on test preparation in the classroom, an emphasis on tested subjects at the expense of non-tested subjects, and a need to better use the data to inform pedagogy (Srikantaiah, 2009). Although teachers have indicated that standards ensure consistency among classrooms, many also report lacking the resources necessary to ensure students are adequately prepared to meet accountability requirements. Additionally, educators have indicated feeling greater
pressure to improve student test scores, especially from administrators, and this impacts morale, contributing to lower job satisfaction and affecting decisions to leave the teaching profession (National Education Association, 2015). Teachers of the most vulnerable student populations are impacted to a greater extent by accountability policies, routinely subjected to administrative oversight. Reback, Rockoff, and Schwartz (2013) determined that inexperienced teachers, in particular, felt anxious about their job security due to the external burdens imposed on them.

The perceptions and attitudes of educators will be explored in this study. Of particular interest is the extent to which teacher attitudes and pedagogical practices have been shaped by policies that have mandated sweeping changes in the evaluation of teaching, often with little regard for the stakeholders who are most impacted (Ballou & Springer, 2015).

Positionality Statement

My interest in this topic stems from my own personal experiences, as well as my professional practice as a teacher and department head. On a personal level, my positionality is influenced by my white, female, middle class background and as a parent of three school-aged children. My professional positionality is shaped by my experiences as a high school Social Studies teacher in three different schools, my interactions with students, parents, and administrators, and my own experience with evaluations.

I was raised in a homogenous community on New Hampshire’s seacoast. My parents instilled in me the value of education from a young age. Both were middle class professionals: my father, a lawyer; my mother, a physical therapist who decided to stay home after I was born. None of my grandparents had attended college so my parents
were the first generation in each of their families to pursue higher education. Coming from working class backgrounds, my parents understood that a college education enabled access to broader economic opportunities. Attending the University of Virginia as an undergraduate student opened my eyes to instances of social inequality. I pledged a sorority during my first semester and joined a Greek community that generally seemed committed to preserving and perpetuating an elite social community. Although I developed some wonderful friendships from the sorority, I never felt like I completely fit into the stereotypical sorority sister mold, and I maintained strong relationships with non-Greek students as well. As a resident of Charlottesville, Virginia, I observed a strong correlation between race and class that had not been evident growing up in New Hampshire. I began to enroll in courses that dealt with topics of social injustice, seeking to learn as much as possible about twentieth century American History in an effort to more fully understand present day inequality. Truly, my time spent at UVA had an enduring impact on my attitudes toward race and class and the formation of my own identity.

As I began a career in education in Massachusetts, my professional experiences reinforced my personal understandings of social inequality. During my first few years in teaching, I was assigned to teach freshmen World History to students who were learning disabled, had behavioral problems, or were English Language Learners (ELL). Recently out of college, where I had focused my studies on 20th century American History, I felt poorly equipped to teach early World History; as it turned out, my shortcomings in content knowledge proved inconsequential. Most of my students struggled with basic literacy and many were on Individualized Education Plans (IEPs). Around this time, a
passing score on the Massachusetts Comprehensive Assessment System (MCAS) exam in reading and mathematics became a graduation requirement; yet most of the students in this class did not have a chance in passing the MCAS, and they knew it. To see these students completely preoccupied with passing the MCAS was disheartening. The focus for general and special education teachers shifted away from helping high-risk students enjoy learning and toward preparing them to pass the MCAS exam. Unfortunately, these objectives were not necessarily one and the same, and as a new teacher, I felt frustrated by the reality of my chosen profession, which was far from my idealized vision.

As I learned more about the backgrounds of my students, I discovered that often they were institutionally, and therefore, academically, disadvantaged. The parents who were knowledgeable about the school system would insist that their children be placed in higher tracks, regardless of ability. Parents of children in the lowest tracks either did not know enough about the school district to advocate for a placement change or deferred to the school for academic decisions. The use of student test scores for assessing teacher quality is something that concerns me due to my experience with teaching students from a variety of ability levels and backgrounds. Teachers, especially those in their first several years, may not have a choice in selecting the courses, levels, or students they teach. These variables, in turn, raise important questions about using student outcomes as a component of teacher evaluations. Students are not uniform. They enter the classroom with different skills, backgrounds, and abilities. Furthermore, their learning is impacted by variables outside of the classroom. The application of a universal standard in the evaluation of teacher effectiveness is complicated by all of these factors.
My personal and professional positionalities continue to overlap and inform one another. I live in a district that is characterized as underperforming, yet am employed in a district that is high performing. I believe in the value of public education, and my children attend public schools; however, I worry that they are being shortchanged, as their teachers seem overly preoccupied with test preparation at the expense of challenging, engaging, and authentic learning experiences. As a new department head, I struggle with implementing evaluation procedures that seem needlessly onerous, and I ponder how to make educator evaluation a worthwhile, reflective, and informative endeavor for both supervisor and teacher.

Purpose Statement and Research Question

The purpose of this study is to examine teacher perceptions as they relate to changes in evaluation procedures. A departure from supervisor observations in favor of student learning data for determining teacher quality led to the overhaul of the system of educator evaluation in Massachusetts. Teachers have been compelled to comply with these changes, with little regard for how the policies impact their profession.

Educators may experience additional pressure from administrators to ensure that their students achieve expected learning benchmarks. Their perceptions are invaluable for understanding how externally imposed mandates affect teaching and learning. The research study will examine the views of participants who are employed as teachers of state-tested subjects in public middle schools in Massachusetts. The research question for this qualitative Interpretative Phenomenological Analysis (IPA) study is:

• What are the experiences of educators whose practices have been shaped by changes to the system of teacher evaluation?
Theoretical Framework

The study will implore the use of an educational change framework to understand how teachers have responded to new evaluation procedures that rate teachers according to performance standards and student achievement. Howley (2012) asserts that organizations undertake change to respond to new political contexts and priorities. This is certainly the case for many schools in Race to the Top states. Teachers’ readiness for change depends on many variables, including buy-in, knowledge, compatibility, and strong leadership. Yet as Khoboli and O’toole (2012) acknowledge, “the teacher is an often forgotten gatekeeper in educational change” (p. 139).

Concerns-based adoption model. The Concerns-Based Adoption Model (CBAM) is appropriate for studying externally motivated, top-down change and how its implementation is perceived among teachers (Gundy & Berger, 2016; Hall & Hord, 1987). Developed by Hall and Loucks (1978), CBAM is based on a set of assumptions:

1. change is a process rather than an event
2. the process is personal
3. successful implementation requires a change in individual classroom practices
4. individuals must change before systems can change
5. teacher change is a developmental process (Horsley & Loucks-Horsley, 1998).

CBAM recognizes that teachers are important stakeholders and their responses to externally imposed change can be understood through an exploration of their perceptions and experiences (Gundy & Berger, 2016). CBAM conceptualizes Stages of Concern (SoC), which represent teachers’ experiences on a continuum that consists of seven
developmental levels: awareness, information, personal concerns, management, consequences, collaboration, and refocusing (Khoboli & O’toole, 2012). Additionally, the Levels of Use (LoU) aspect is designed to gauge teachers’ confidence and abilities in the implementation of the change. Teachers’ placement on the SoC continuum will impact their LoU, which will affect their perceptions of the policy changes.

According to Galley (2011), “An effective teacher evaluation system requires educator buy-in, research-supported measures, and resources to support implementation of the plan” (p. 44). Educators’ feelings about the adoption of changes to evaluation is contingent on school leaders’ abilities to communicate reasons for the change, establish buy-in, and work with teachers to craft a shared vision for new behaviors. The Concerns-Based Adoption Model is a suitable analytical framework for this study since public school teachers in Massachusetts have had to adjust to a different set of expectations due to the implementation of state-mandated revisions to the evaluation system during the past few years. Stages of Concern and Levels of Use will be utilized to explain how teachers have navigated the demands placed upon them due to these changes.

A review of the literature on the reliance of growth models for teacher evaluation reveals that while there are many studies that examine the instruments used for assessing student learning, as well as the impact of high-stakes testing on the curriculum, teaching, and learning, few studies take into account teachers’ perceptions of policies that link student achievement measures with teacher evaluation. The researcher will seek to understand how educators experience teacher evaluation methods that link teacher effectiveness with student achievement.
Chapter 2: Literature Review

The expanding role of the federal government in the realm of public education has contributed to policies that impact teacher evaluation. In Massachusetts, recently revised educator evaluation requirements have been implemented in school districts across the commonwealth, and teachers have been compelled to adapt to these changes. In most cases, educators have not had a voice in the implementation of these policies, and therefore, this study aims to bring to light the experiences of teachers in order to inform district leaders and policymakers of the effect on the teaching profession. The purpose of this chapter is to provide a review of the relevant literature on teacher evaluation methods and the increasing use of student achievement data in the assessment of teacher effectiveness.

The literature review will seek to trace the evolution of standards-based reforms over the past several decades, revealing how greater centralization has led to policies that affect public educators, who have had to adapt their practices to comply with teacher evaluation requirements, often with little regard for their professional judgment (Everson et al., 2013; Lee, 2011). Therefore, the theoretical lens provided by the Concerns-Based Adoption Model will undergird the study, framing an investigation into how teachers have responded to changes in evaluation protocols. Whereas teachers have had little input regarding the implementation of federal and state mandates, the study is aligned with a constructivist paradigm in an effort to illuminate the experiences of educators whose professional lives have been altered as a consequence of these policies. The literature review is organized to provide an historical overview of teacher evaluation
systems, describe and explain the purpose of student growth models, and explore the educational impact of increased teacher accountability in terms of student achievement.

**Historical Perspective on Teacher Evaluation Systems**

It is essential to examine the history of educational reform movements at the national level over the past several decades in order to better understand the educator experience from an historical perspective. Although standardization and accountability have been synonymous with educational reform since World War II, when government policies began to rely on tests and assessments in some capacity to provide information about the nation’s public schools, student performance did not become a factor in individual teacher evaluation until recently, when the Obama administration launched the Race to the Top competitive grant program to encourage states to hold teachers more accountable for student learning in evaluation systems (Ballou & Springer, 2015; Goldhaber, DeArmond, & DeBurgomaster, 2011).

Previously, each educational reform movement utilized testing in different ways (Linn, 2000). The first widespread testing initiative occurred during the 1970s, due to concerns about America’s position internationally, as outsourcing of manufacturing jobs and advancements in technology instilled fear in many policymakers and educational reformers that Americans were falling behind their global counterparts (Amrein & Berliner, 2002). The persistent belief that American students lacked the adequate skills to compete in the global economy provided the rationale for the minimum competency movement, which represented a shift in the national educational objectives (Hursh, 2005). The minimum competency movement signaled an increasing reliance on standardized tests and student data measurements to ensure that students achieved minimum
benchmarks for learning (Amrein & Berliner, 2002). The data was also used to assess schools and for decisions related to promotion, retention, graduation, and tracking of students. During the 1980s and 1990s, state reliance on competency testing grew, with thirty-three states imposing minimum competency tests by the mid-1980s and eighteen enforcing test-based requirements for graduation by the mid-1990s (Heubert & Hauser, 1998).

The 1983 publication *A Nation At Risk* set the tone for the contemporary standards movement by asserting that “the educational foundations of our society are presently being eroded by a rising tide of mediocrity that threatens our very future as a Nation and a people” (National Commission on Excellence in Education, 1983, p. 5). In the aftermath of the report, states established commissions to examine educational outcomes, with many instituting reforms that promoted more testing and increased curricular requirements for students (Au, 2009). Ultimately, *A Nation At Risk* bolstered political and public support for higher standards and increased accountability measures, despite educators’ concerns with using test scores as an incentive for raising student achievement (Hollingworth, 2009).

**No Child Left Behind (NCLB).** The proliferation of standardized testing continued throughout the 1980s and 1990s, as schools attempted to demonstrate they were improving student outcomes (Amrein & Berliner, 2002; Heubert & Hauser, 1998). The educational reforms of the twenty-first century are rooted in the notion that the quality of education in American schools varies considerably due to inequality. Inequality in education became a national issue during the 1960s, when efforts to desegregate schools exposed vast discrepancies in the quality of education that existed
due to traditions of state and local control (Hollingworth, 2009; Linn, 2000).

Hollingworth (2009) explains that the passage of the Elementary and Secondary Education Act (ESEA) of 1965 increased federal involvement in public education by providing federal funding for schools in an attempt to increase access to quality education for poor and minority students, thereby promoting social mobility. Whereas the ESEA allocated need-based aid, No Child Left Behind, which was in fact a reauthorization of the ESEA, provided need-blind aid, facilitating a more encompassing federal role in public education (Hollingworth, 2009). Under NCLB, states and school districts were required to track and report on students’ academic progress, and states had to prove that students were making Adequate Yearly Progress (AYP) through annual assessment measures (Pearson Education, Inc., 2004; Toutkoushian & Curtis, 2005). NCLB also called for teachers to be “highly qualified,” emphasizing content knowledge over pedagogical skills. The legislation allowed states to determine teacher qualification requirements, but in general teachers are defined as highly qualified if they meet state licensing requirements, pass a rigorous content-specific test, and hold a Bachelor’s degree or higher (Smith & Gorard, 2007).

Over time, NCLB’s requirements became increasingly burdensome for school districts, and critics claimed that the legislation could never realize its intended objectives if student achievement was measured using standardized tests (Gay, 2007). In December 2015, NCLB was effectively eclipsed by the Every Student Succeeds Act, a reauthorization of the Elementary and Secondary Act of 1965, which once again reaffirmed the nation’s commitment to the promotion of equal opportunities for all students (Every Student Succeeds Act, 2017).
**Race to the Top (RTTT).** Although no longer relevant, NCLB set into motion a series of national policies that emphasized outcomes above all else. In terms of teaching, NCLB focused on training and qualifications, while the American Recovery and Reinvestment Act of 2009 underscored teachers’ abilities to improve student achievement (Mangiante, 2011). Often referred to as Race to the Top (RTTT), this federal initiative committed millions of dollars in federal support to states that adopted measures to more accurately determine the value a teacher adds to student learning from year to year (Amrein-Beardsley & Collins, 2012). RTTT effectively shifted attention away from individual school performance and toward individual teacher performance. The demands for increased accountability coincided with a focus within the research community on identifying and proliferating quality teachers (Schacter & Thum, 2004). This shift in policy extends high-stakes tests to include teachers, who now have a vested interest in ensuring students perform well on assessments since their own evaluations may be linked to student outcomes.

**Common Core State Standards (CCSS).** As a condition of the 2009 Race to the Top (RTTT) initiative, states had to agree to adopt standards and assessments “that prepare students to succeed in college and the workplace and to compete in the global economy” (U.S. Department of Education, 2015). The Common Core State Standards were developed in 2009 by state leaders from forty-eight states, two territories, and the District of Columbia “to ensure all students, regardless of where they live, are graduating high school prepared for college, career, and life” (Common Core State Standards Initiative, 2014). The Common Core standards are intended to provide a clear set of national learning expectations for the knowledge and skills of students at each grade level
in English Language Arts (ELA) and Mathematics, yet the implementation of Common Core standards is reserved for the states and local school districts. Although the Common Core does not dictate how educators should teach, many states decided to consult with the Partnership for Assessment of Readiness for College and Careers (PARCC) or the Smarter Balanced Assessment Consortium (Smarter Balanced), which are consortia that worked to develop assessments to inform teaching and evaluate learning. Presently, forty-two states, four territories, the District of Columbia, and the Department of Defense Education Activity (DoDEA) have adopted the Common Core and are in the process of implementing the standards (Common Core State Standards Initiative, 2017).

The widespread adoption of the Common Core State Standards in the United States continues to shift public school control away from the local community and toward the state and federal government, while at the same time imposing more conditions on schools, administrators, and teachers (Guarino, Reckase, & Wooldridge, 2015). Testing companies stand to profit from the greater reliance on newly developed assessments designed to align with the Common Core, and although the alleged goal of Common Core is to prepare students for college and life in the 21st century, the overreliance on high-stakes assessments ultimately leads to a narrowing of the curriculum in most subjects (Karp, 2013). Jones (2007) found that the preoccupation of many school districts with student test performance limits the autonomy of teachers in their classrooms and forces them into a role where they are expected to merely develop basic academic skills at the expense of creativity, innovation, problem-solving, and interpersonal skills. Teachers report feeling additional pressure from building and district administrators to
improve standardized test scores and express frustration with necessary drill and skill exercises, which tend to be less enjoyable for students, at the expense of content development (Agee, 2004; Jones, 2007).

In Massachusetts, revised teacher evaluation requirements equate teacher performance ratings with student performance data. Teachers must align their content to common assessments (formerly called District-Determined Measures), which will be used to measure student growth and assess the value added by the teacher (Massachusetts Department of Elementary and Secondary Education, 2017). States that received federal funding under the Race to the Top program were compelled to revise their teacher evaluation systems as a condition of aid. Criticism over the lack of rigor in traditional teacher observations and concerns over the “widget effect,” a report that exposed difficulties in distinguishing good teachers from bad, further reinforced the belief among policymakers that evaluation should reflect performance (Weisberg, Sexton, Mulhern, & Keeling, 2009). As a result, new models of teacher evaluation that incorporate student-learning data have been adopted by many states. Under the Race to the Top program, nineteen states, including Massachusetts, reformed teacher evaluation methods to more accurately reflect teacher quality (Hallgren, James-Burdumy, & Perez-Johnson, 2014). In the era of accountability, school districts have been under increasing pressure to emphasize student-learning gains as the main predictors of teacher effectiveness, rather than the more observational qualities of personality and relatability. These new measures of teacher effectiveness have generated controversy, however, due to concerns over the reliability of their accuracy. In addition, teachers may be assessed based on student test
scores but have no idea how to use that data to improve their teaching (Amrein-Beardsley & Collins, 2012; Collins, 2012).

While formal teacher evaluation processes are nearly a century old, only recently have the methods used for evaluating teachers come under scrutiny for not aligning with student learning outcomes (Attinello, Lare, & Waters, 2006; Strong et al., 2011). A majority of teachers believe that evaluation should be used as a tool for guiding professional development and promoting reflection, but too often neither of these outcomes result from formal evaluation practices (Peterson & Comeaux, 1990). Wise, Darling-Hammond, McLaughlin, & Bernstein (1985) assert that “the design of teacher evaluation systems depends critically on educational goals” (p. 67). In theory, effective teacher evaluation methods should generate self-efficacy among faculty and lead to improved pedagogy (Bandura, 1982). Conventional methods of teacher evaluation, however, typically assess educators on the basis of licensing requirements, qualifying examinations, and ratings from classroom observations, which may or may not correlate with student achievement (Soar, Medley, & Coker, 1983). As a result, traditional teacher evaluations based mostly on observations are often viewed as an exercise that has little impact on practice or instruction (Attinello et al., 2006). The high percentage of teachers who are rated proficient under this system, combined with the prevailing assumption that the teaching profession is infested with poorly qualified teachers, has caused researchers to question whether subjective evaluations can actually serve as accurate measures for assessing teaching effectiveness due to the potential for bias (Ambady & Rosenthal, 1993; Jacob & Lefgren, 2008; Medley & Coker, 1987; Strong et al., 2011).
**Current evaluation system.** The framework for current teacher supervision emerged in the 1960s and 1970s as a process of reflective dialogue between teachers and their supervisors that consisted of a pre-conference, classroom observation, analysis, post-conference, and analysis of the analysis (Marzano, Frontier, & Livingston, 2011). One of the major criticisms of traditional evaluation methods is that supervisors’ opinions are preconceived and therefore unreliable (Jacob & Lefgren, 2008; Strong et al., 2011). Studies have investigated whether principals’ assessments of teacher effectiveness were based on the teachers’ impact on student learning or some other characteristic (Ambady & Rosenthal, 1993; Medley & Coker, 1987; Strong et al., 2011). Medley and Coker (1987) found that there was little correlation between student learning gains and principals’ perceptions of teachers, concluding that principals’ ratings of teacher effectiveness measure some other desirable teacher characteristic. Ambady and Rosenthal (1993) examined strangers’ perceptions of teachers based on “thin slices” of the individuals’ behavior and determined there was an association between the strangers’ and principal’s perception of the teachers, which confirmed the importance of teacher affect. Both of these studies revealed that likeability is a desirable quality in teachers, although this characteristic is not necessarily equated with greater student learning outcomes.

Indeed, research indicates that observations of teacher quality do not necessarily correlate with data obtained from student achievement indicators (Jacob & Lefgren, 2008; Strong et al. 2011). Strong et al. (2011) also utilized “thin slicing” in their study, gathering the opinions of one hundred judges from ten different categories of education and non-education backgrounds. The researchers compared the results from the judges’
analyses of perceived teacher effectiveness with calculated value-added scores from district student achievement data. They concluded that the judges achieved “relatively high levels of agreement but were absolutely inaccurate” (p. 378) when accounting for value-added measures, calling into question the reliability of supervisors’ observations of teacher effectiveness. Jacob and Lefgren (2008) examined whether principals could distinguish between more and less effective teachers where effectiveness is defined by one’s ability to raise student scores in math and reading, discovering that principals typically base their opinions of teachers on formal and informal observations, reports from parents, and student achievement scores, but the degree to which each of these components influences a principal’s assessment of teachers varies. In comparing principals’ ratings of teachers with student achievement data, the research concluded that principals do effectively identify the best and worst teachers among their staff, but they have difficulty distinguishing among those in the middle (Jacob & Lefgren, 2008). The inability of principals to classify teachers according to effectiveness is problematic for high-stakes personnel decisions, such as merit pay, remediation, promotion, or dismissal. Furthermore, observations in isolation do not provide school leaders with specific diagnostic information to create meaningful professional development opportunities for enhancing teaching and learning.

An examination of the historical trends in evaluation is essential for fully appreciating how this evolution has influenced the experiences of teachers. Whereas education was once viewed as a local responsibility, community control has diminished as a result of national legislation and federal mandates. The economic uncertainty of the 1970s, combined with a pervasive fear that America was falling behind other nations,
contributed to the increasing use of standardized tests in schools and more curricular requirements for students. Policymakers expected higher educational standards to translate into better economic outcomes. NCLB further solidified political and public support for standards and accountability, requiring states and school districts to report on students’ academic progress and establishing greater credentialing requirements for public school teachers. RTTT effectively shifted the focus away from student achievement and toward the value an individual teacher adds to learning, while simultaneously providing an opportunity to make evaluation less subjective. Regardless of their credentials, positive feedback from administrators, parents, or students, or their ability to instill a love of learning in their students, teachers employed in states that received RTTT funding are deemed effective primarily based on student learning gains. Subsequently, the next section will discuss some of the most commonly used student growth models and their purpose in educator evaluation. It is necessary to examine these models in order to fully inform the current experiences of public school teachers.

**Student Growth Models**

Increased standardization, as well as efforts to promote educational equity through improved teacher quality, contributed to the development of methods for evaluating teachers more objectively, using a tool known as a student growth model. Reliance on student growth models set into motion a process for assigning student achievement data to specific educators. The momentum behind student growth models occurred as a natural extension of No Child Left Behind (Stronge, Ward, Tucker, & Hindman, 2007). Under NCLB, school districts were required to demonstrate that they were making AYP, as determined by student performance on statewide tests. Analysis of
AYP was typically made using a cohort comparison, comparing the student test scores of the current year with the student test scores of the previous year. Critics of the cohort comparison model assert that level indicators, which determine the quality of a school based on the average achievement of the students, “produce misleading assessments of schools with large numbers of academically disadvantaged students” (Pearson Education, Inc., 2004, p. 6) and divert teacher focus away from the highest and lowest achievers in favor of those students who will make it possible for the school to reach AYP. Proponents of student growth models claim that these tools are a more accurate predictor of student growth and school progress than cohort comparisons since they follow individual student achievement over time (Karl, Yang, Lohr, 2013; Scherrer, 2012). 

Lefgren and Sims (2012) suggest that “removal of the lowest performing teachers, even using imperfect value-added metrics, will lead to important gains in student learning over time” (p.109). A general consensus on this topic among policymakers led at least forty states to use, pilot, or develop growth or value-added models as a component of teacher evaluation (Collins, 2012). Whereas growth models typically focus on the performance of individual students, value-added models address the effects of individual teachers on student achievement gains. Although these models differ in scope, they each provide information about student learning and its relationship to teaching. Both growth and value-added models are preferable to status models, which correlate more with non-school factors, such as community characteristics, than school factors, such as resources, policies, and leadership (Franco & Seidel, 2014). As opposed to status models, which have often been criticized for rewarding teachers based on whom they teach since they measure student performance from a single test at a snapshot in time, value-added models
aim to attribute student achievement gains to individual teachers (Scherrer, 2012). The requirement under RTTT that states adopt measures to rate teacher effectiveness has hastened the widespread adoption of student growth models as components of teacher evaluation. As a result, teachers are more accountable for student learning gains, particularly those who teach in subjects where standardized tests are available and administered.

**Value-added models.** The introduction of one of the most common growth indicators, known as the value-added model (VAM), a statistical method designed to evaluate student growth and estimate teacher effectiveness, presented schools with an instrument for measuring student growth in subjects such as reading and math and assessing the added “teacher effect” on student learning over the course of the school year (Ballou, Sanders, & Wright, 2004; Braun, 2005; Darling-Hammond et al., 2012; Doran & Fleischman, 2005; Hanushek, Kain, O’Brien, & Rivkin, 2005; Hill, Kapitula, & Umland, 2011; McCaffrey, Lockwood, Koretz, & Hamilton, 2003; Yeh, 2012). Consensus within the research community on the importance of teachers in student learning has contributed to the proliferation of standardized tests designed to measure student achievement and attribute the growth to individual teachers (Sanders & Horn, 1998; Sanders & Rivers, 1996). Policymakers and researchers have increasingly turned toward statistical measures for evaluating teachers due to the persistent belief that observational evaluations are largely subjective and do not necessarily reflect the teacher’s impact on student learning (Jacob & Lefgren, 2008; Strong et al., 2011). The adoption of VAMs for teacher evaluation and high-stakes decisions for compensation,
and even employment, has resulted in considerable research on the reliability of such models.

**Tennessee value-added assessment system.** One of the earliest value-added models was adopted by William Sanders, a statistician and chief architect of the Tennessee Value-Added Assessment System (TVAAS). Sanders developed “an efficient and effective method for determining individual teachers’ influence on the rate of academic growth for student populations” (Sanders & Rivers, 1996, p. 1) in response to Tennessee’s Education Improvement Act of 1991, which called for greater accountability and assessment measures to improve education. The TVAAS revealed that the teacher is the most important factor in student achievement (Sanders & Horn, 1998). Whereas highly effective teachers could elevate student learning, ineffective teachers had the opposite effect. In fact, students who had ineffective teachers experienced residual effects for several years after (Sanders & Rivers, 1996). Educational inequity is compounded when students are assigned to ineffective teachers year after year. Early findings of the TVAAS suggested that if administrators were able to determine teacher effectiveness using data from student test scores, they would be capable of deliberately assigning the most needy students to the most effective teachers, thus leveling the playing field for students who were systematically disadvantaged. In addition, the results could provide administrators and teachers with formative data that could be instructive in the design of professional development for less effective teachers.

**Education value-added assessment system.** In 2000, Dr. Sanders took the TVAAS to the SAS Institute, Inc., an analytical software company whose goal is to aid consumers in making the best business decisions possible, where he changed the name to
Education Value-Added Assessment System (EVAAS) (SAS, 2013). In 2003, the RAND Corporation released a report evaluating value-added models. The researchers concluded that value-added models had attracted increasing attention for at least two main reasons: 1) VAMs had the potential to separate the effects of teachers and schools from other noneducational factors, such as family background; 2) early VAM studies appeared to show large differences in teacher effectiveness (McCaffrey et al., 2003). Value-added models gained a reputation for identifying effective and ineffective teachers, regardless of external factors on student learning (Hershberg, Simon, & Lea-Kruger, 2004). Claiming to comply with federal legislation that called for greater school and teacher accountability, the EVAAS, which was marketed to school districts nationwide, was piloted in 15 states between 2005 and 2010 (Collins, 2012).

In the evaluation of Sanders’ EVAAS, Hershberg et al. (2004) conceive that it is possible to discern whether or not instruction was effective by comparing students’ end of the year scores with their projected values. Furthermore, their assessment indicates that statistically advanced VAMs, like the EVAAS, account for socioeconomic differences since students serve as their own controls. Sanders and Horn (1998) claim that value-added models provide “the effective and appropriate use of standardized test data for the purpose of educational evaluation” (p. 255). Their findings suggest that VAMs are the most accurate predictors of individual teacher effectiveness.

**Improving teacher quality.** Researchers who support using student achievement data as a component for teacher evaluation recognize potential for reforming teacher tenure and removing ineffective teachers from the classroom (Hanushek, 2009; Harris, 2009; Rivkin, 2009; Winters & Cowen, 2013).
**Teacher deselection.** Hanushek (2009) asserts that the least effective teachers are harming the profession and damaging students’ educational experiences. Emphasizing the relationship between teacher quality and student achievement, he attributes declining international prestige to the persistence of poor quality teachers, who may be difficult to remove from their positions due to teacher tenure policies. He recommends policies that reevaluate teacher tenure and ultimately allow for the “deselection” of ineffective teachers, suggesting that to do so would result in an overall gain in GDP, which would cover the increased costs to K-12 education in terms of salary adjustments and removal procedures. Rivkin (2009) concurs, but acknowledges that the basis for determining teacher quality must be perceived as “informative and fair” in order to be considered valid and avoid opposition from teachers and unions. Winters and Cowen (2013) also support reforming teacher tenure, encouraging value-added measures to deselect teachers based on consecutive poor performance scores as a strategy for improving long term teacher quality.

**Compensation reform.** Researchers differ in their views on connecting student achievement gains to educator compensation. Some researchers contend that teacher compensation should be based partially on an educator’s ability to demonstrate an impact on student learning, and student growth models provide administrators with additional insight for linking performance with pay (Rivkin, 2009). Innovative approaches to educator pay could satisfy those analysts and policymakers who often criticize teacher compensation systems “for being too uniform and rigid” (Goldhaber et al., 2011, p. 441), rewarding teachers for years of experience and credentials, such as degrees earned. The
salary schedules utilized by the overwhelming majority of school districts pay teachers according to characteristics that are weakly connected to teacher quality.

Critics assert that strict reliance on salary schedules rewards those of less quality who are already in the profession with pay increases for years of service, while potentially discouraging high-achieving individuals from entering the teaching profession since there is little financial incentive, particularly for those with advanced technical skills (Goldhaber et al., 2011). In other professions, such as medicine and higher education, pay differentials may be due to productivity or specialty, and this flexibility allows for “greater overall cost effectiveness” (Podgursky & Springer, 2011, p. 168). In the public K-12 system, however, teachers are compensated according to experience and education, and neither of these factors are necessarily predictors of teacher quality (Hanushek, Rivkin, Rothstein, & Podgursky, 2004; Podgursky & Springer, 2011). Schacter and Thum (2004) encourage school districts to implement merit pay incentives for teachers, suggesting that the most significant concern with the current system is that “no teacher’s job or career security is dependent on performance” (p. 420).

Value-added measures provide school districts with an opportunity to evaluate the contributions of individual teachers, while simultaneously reforming teacher compensation systems by rewarding those identified as “effective” with bonuses (Rivkin, 2009). This could reduce or eliminate the obligation of school districts to pay higher salaries to less effective teachers simply because they have advanced degrees and more years of service. Some education reformers propose using incentive programs for educators, which could serve to elevate the status of the teaching profession by attracting and retaining higher-quality individuals and weeding out less effective teachers
Underscoring this theory is the belief that there is not only an educational advantage for adopting student growth models, but also an economic advantage; retain and reward high-quality teachers, demote and remove costly low-quality teachers.

The increased costs associated with the implementation of student growth models may be particularly burdensome for school districts, especially if the results are deemed inaccurate. Yeh and Ritter (2009) concluded that rapid assessment, the process of testing students two to five times a week in math and reading, was a more cost-effective alternative for raising student achievement when compared to other methods, including replacing the bottom quartile of teachers identified through value-added modeling. Additional research on the cost-effectiveness of replacing the bottom 40% of teachers demonstrated that the net effect of this strategy would be to reduce student achievement and result in additional costs, mostly due to the “high probability that the use of VAM to terminate teachers would be challenged in almost every case, resulting in high legal costs” (Yeh, 2012, p. 394). While the research on the economic costs and benefits for student growth models is conflicting, the results indicate that since such models may lack reliability, they should not be used for compensation decisions.

Teacher evaluation and student achievement, once examined in isolation, have become inextricably linked as a result of the RTTT competitive grant program, which compelled recipient states to revise their teacher evaluation systems to incorporate student growth measures. The apparent success of some value added models, such as the TVAAS and EVAAS, which purportedly could be used to determine student achievement gains regardless of student background and other non-educational factors, led to the
widespread adoption of VAMs in evaluation. Proponents of VAMs claim that these instruments allow districts to dismiss poor performing teachers, improving overall teacher quality, while also enabling compensation reform. Yet the research indicates that student growth models may prove more costly than cost effective, as teachers who dispute the accuracy of the results may create a financial burden for districts compelled to defend the legality of their actions if challenged by unions (Yeh, 2012). Significantly, the implementation of new evaluation systems extends beyond improving teacher quality and more carefully aligning compensation with excellence. Indeed, the implementation of new evaluation methods that consider student achievement has far reaching educational consequences. The next section will discuss how student growth models are potentially problematic for determining teacher quality and address how testing affects curriculum, teaching, and learning.

**Educational Impact of Value-Added Teacher Accountability**

The haste with which new accountability measures for teachers have been adopted is cause for alarm (Darling-Hammond et al., 2012). Opportunities for testing companies to profit or for school districts to reduce operating costs by dismissing highly paid teachers with low student impact ratings expose the potential economic incentives for championing value-added models (Karp, 2013; Rivkin, 2009). While most researchers and practitioners agree that value-added assessments may offer benefits for analyzing student learning and improving teacher instruction, quite a few studies have uncovered potential problems with these tools, highlighting the weaknesses and unintended consequences of reliance on such measures for determining teacher
effectiveness (Darling-Hammond et al., 2012; David, 2010; Papay; 2011; Rothstein, 2010; Schafer, Lissitz, Zhu, Zhang, Hou, & Li, 2012).

**Different tests.** As more school districts across the nation use student growth models to measure student growth and assess the contributions of individual teachers, researchers need to evaluate whether various methods serve the same function (David, 2010; Papay, 2011). While the EVAAS model has been one of the most widely adopted value-added models, teachers and administrators have expressed confusion over interpreting the data (Collins, 2012). The EVAAS uses complex statistical mixed models and covariant structures to project and measure actual student growth. While most education professionals agree that VAM data should be used formatively, few seem to understand how to do so (Amrein-Beardsley & Collins, 2012). Hershberg et al. (2004) do not view this as a problem: “Not everyone who uses a personal computer or a television remote needs to understand the intricacies of micro engineering; similarly, not everyone who uses value-added assessment needs to understand its complex methodology” (p. 28).

Some studies suggest that simple value-added models are as effective as more statistically complex versions (Doran & Fleischman, 2005; Lefgren & Sims, 2012). Lefgren and Sims (2012) developed a simple value-added model to be used in elementary schools, where teachers are often responsible for teaching several subjects. Their results demonstrate that school districts could readily implement low-cost “methods of improving the predictive power of value-added measures” (Lefgren & Sims, 2012, p. 120) by relying on the subject test scores. Sanders and Wright (2008) maintain that the results from simplistic VAMs “may over-identify either very ineffective or very effective
teachers” (p. 6). Papay (2011) concurs that different tests produce “substantially different estimates of individual teacher effectiveness” (p. 166). Additionally, Papay cautions policymakers against using VAMs for high-stakes decisions about teacher performance since “all value-added estimates of teacher effectiveness use tests designed to measure student, not teacher, performance” (p. 189). This is an important factor for not only policymakers, but also administrators, to consider when placing an emphasis on VAMs for determining teacher effectiveness.

**Inconsistent results for teacher quality.**

I do what I do every year. I teach the way I teach every year. [My] first year got me pats on the back. [My] second year got me kicked in the backside. And for year three my scores were off the charts. I got a huge bonus, and now I am in the top quartile of all the English teachers. What did I do differently? I have no clue. (Amrein-Beardsley & Collins, 2012, p. 15).

This statement from an eighth grade advanced English teacher in the Houston Independent School District (HISD) highlights the confusion and concerns of researchers and practitioners with the increasing reliance on value-added models for teacher accountability and evaluation. A significant problem with the use of student growth models for teacher evaluation is the lack of research that validates their reliability for determining teacher quality (Amrein-Beardsley, 2008). A recent mixed method study found that, in some instances, teachers who received VAM scores in the first and second quintile presented material in a way that “appeared potentially harmful to students’ learning” (Hill et al., 2011, p. 824). Another study determined that policies based on
value-added assessments would reward and punish the wrong teachers (Rothstein, 2010). Even more troubling, teachers do not have a clear understanding of how to modify their instructional practices to improve student learning, and in turn, their own VAM scores (Amrein-Beardsley & Collins, 2012; Hill et al., 2011). Additionally, research reveals that student achievement gains are due to multiple factors, not just the contribution of an individual teacher; therefore, the assumption “that student learning is measured well by a given test, is influenced by the teacher alone, and is independent from the growth of classmates and other aspects of the classroom context” (Darling-Hammond et al., 2012, p. 8) is flawed. Many students may work with academic support personnel, tutors, or other instructional leaders who also influence their learning (Schafer et al., 2012). Thus, while VAMs may serve a function in education, these models are unreliable as tools for teacher evaluation. David (2010) recommends using multiple measures to evaluate teachers and protect them “from erroneous and harmful judgments” (p. 82) that might result from overdependence on VAMs.

**Student sorting, student characteristics, and missing data.** Perhaps the most serious issue with value-added models results from claims that VAMs “provide a means of separating the effects of schools and teachers from those of such noneducational factors as student background” (McCaffrey et al., 2003, p. 111). In theory, since students serve as their own controls, many VAMs do not account for student background factors. The efficacy of VAMs for comparing teachers within schools, and schools with other schools, however, rests on the assumption that students are assigned randomly to teachers. Rothstein (2009) asserts that student assignments are often not random and this complicates the VAM data. Principals may group students of similar ability together, or
they may assign students with a range of abilities across classrooms. Some teachers who have earned a reputation for teaching a particular skill may be assigned students who are in need of additional work in that particular area, while others may receive the easiest to teach students as a special favor from the administrator. Additionally, parents may request certain teachers for their children. All of these unobserved variables influence classroom assignments and bias estimates of teachers’ effects (Koedel & Betts, 2011; Rothstein, 2009; Steele, Kraemer & Meyer, 2012).

Rothstein’s (2009) research reveals that “a teacher assigned students with high 4th grade gains in the previous year will look like a bad teacher through no fault of her own, while a teacher whose students posted poor gains in the previous year will be credited for their predictable reversion to trend” (p. 3). Clotfelter, Ladd, and Vigdor (2006) found that teachers “with more experience, degrees from more competitive colleges, and advanced degrees tend to teach at schools with more affluent, higher achieving and whiter populations” (p. 779). Nonrandom sorting also occurs within schools, where educators with stronger credentials teach the most affluent students and those with less experience are “assigned less able students” (p. 782). These findings reveal that efforts to isolate teacher effectiveness using VAMs are complicated by the nonrandom sorting of students geographically and within schools. Therefore, administrators should exercise caution when using student test scores to compare teachers unless random sorting of students and heterogeneous grouping is ensured.

The most widely utilized value-added models do not account for differences in student background. Therefore, teachers of students who are English Language Learners (ELL), or have learning disabilities, behavioral problems, or poor attendance may be
disadvantaged by their student assignments (Amrein-Beardsley, 2009; Darling-Hammond et al., 2012). Teachers of high-achieving gifted students may also be disadvantaged since exceptional students often post the least significant gains over the course of the school year (Amrein-Bearsley & Collins, 2012). Covariates, confounding factors, and missing data all impact the calculation of a teacher effect (Lomax & Kuenzi, 2012). Covariates are those factors that may affect a student’s academic achievement, but are not attributed to the teacher, such as socioeconomic status, disability, and ELL status. Confounding factors are those “within the culture of the school, community, or neighborhood that can influence the teacher effect” (Lomax & Kuenzi, 2012, p. 8). Even though VAMs claim to reduce the influence of confounding factors, they cannot be eliminated entirely, and this makes it difficult to compare teacher effects across schools, districts, or states. Additionally, missing data due to student absences or mobility can bias VAM data, either positively or negatively depending on the data this is missing (Anderman, Anderman, Yough & Gimbert, 2010; Lomax & Kuenzi, 2012). These factors highlight problems with value-added assessments and indicate a need for further research prior to implementation for reward or remediation.

**Educational consequences for curriculum, teaching, and learning.** Proponents of standardization and accountability claim that increased reliance on tests to determine student learning gains and evaluate teachers will ensure that students across America receive a basic level of education, regardless of the community in which they live, the school they attend, their class, race, or gender (U.S. Department of Education, 2016). Supporters of increased reliance on assessments assert that improved test scores and measurable student achievement will generate feelings of success among teachers and
students, whereas low test scores will encourage teachers to revise their lessons and instill in students a stronger determination to learn (Amrein & Berliner, 2002). Furthermore, poor student performance will reflect back on the teacher and allow administrators to base employment decisions and professional development offerings on test results (Harris, 2009). These beliefs assume that tests are necessarily effective measures of content taught to students, provide an equal opportunity for students to demonstrate their knowledge and skills, and “are good measures of an individual’s performance, little affected by differences in students’ motivation, emotionality, language, and social status” (Amrein & Berliner, 2002, p. 5). Numerous studies on the effects of high-stakes testing on curriculum, teaching, and learning contradict the claims made by proponents (Agee, 2004; Amrein-Berliner, 2003; Costigan, 2002; Darling-Hammond, 2007; Diamond & Spillane, 2004; Hollingworth, 2009; Horn, 2003; Jones, 2007; Polesel et al., 2012; Scherrer, 2012). Truly, high-stakes tests often have unintended consequences that are detrimental to the system of public education in general and to specific student populations in particular. Tying teacher evaluation to student test scores only intensifies these concerns.

**Impact on curriculum.** The emphasis on student tests scores in state and federal policy has had a significant impact on the design of curricula in schools (Diamond & Spillane, 2004; Grant, 2004; Hollingworth, 2009; Jones, 2007; Madaus & Clarke, 2001; Polesel et al., 2012). Schools have had to modify course material and offerings in order to ensure that students have learned the content and skills on which they will be evaluated (Grant, 2004). According to Diamond and Spillane (2004), high-stakes testing that attaches rewards and punishments has prompted administrators and teachers to develop
curricula focused on raising test scores. In some states, mandatory subject testing forced schools to realign curriculum to satisfy testing requirements, with many schools now devoting considerably more instructional time to subjects that have high-stakes tests attached, like mathematics and English, resulting in less instructional time spent on other subjects (Jones, 2007). Since the implementation of NCLB, the curriculum in many public schools, especially those labeled as low performing, has been structured around the content of standardized tests (Polesel et al., 2012).

Intense concentration on test preparation due to external pressures has resulted in a narrowing of the curriculum, which denies students a diverse and enriching educational experience (Au, 2007). The way in which schools respond to high-stakes accountability policies is situated within the school’s status, with probation schools focusing on narrowly complying with policy demands and higher performing schools seeking to enhance the performance of all students regardless of grade level and across all subjects (Diamond & Spillane, 2004). In public elementary schools across the country, less time is devoted to science and social studies, and some schools have eliminated music and art programs in order to allow for more time on tested subjects. The reduction of arts related curricula is particularly unfortunate considering students with high levels of arts education have been found to outperform students with low levels of arts education on standardized assessments, regardless of socioeconomic background (Catterall, Chapleau & Iwanaga, 1999, as cited in Polesel et al., 2012). Schools with limited financial resources have less flexibility when it comes to offering enriching curricular experiences that extend beyond the core curriculum; therefore, students of low socioeconomic status have more to lose compared to those of middle-class and affluent backgrounds when
curriculum is designed around tested subjects (Cunningham & Sanzo, 2002; Diamond & Spillane, 2004).

**Impact on teaching.** The policies of high-stakes testing have had a primarily negative effect on the teaching profession (Grant, 2004; Madaus & Clarke, 2001; Polesel et al., 2012). In the early 1900s, educational reformer John Dewey proposed that learning be student centered, authentic, and based on real world experiences which reduce the isolation of the school; however, one hundred years later, high-stakes testing policies encourage teacher-directed drills and repetition (Dewey, 2001). Madaus & Clarke (2001) found that when “teachers perceive that important decisions are related to the test results they will teach to the test” (p. 6). Due to the increased state and federal accountability measures, teachers feel additional pressure from building and district administrators to improve standardized test scores. This anxiety increases when teachers report a higher percentage of minority students in their classes (Madaus & Clarke, 2001). Teachers are under pressure to cover all the content on the tests, and this leads to a direct instruction classroom environment, which typically fails to account for diverse student learning styles (Polesel et al., 2012).

Another consideration of high-stakes testing policies is that teachers may not have the freedom to teach content in a way they feel would enhance student learning if they must focus on preparing students for a specific test. Agee (2004) discovered that a young English teacher had to abandon many of her original goals for teaching multicultural literature in order to adequately cover the material on which her students would be tested. New teachers may be especially vulnerable if teacher preparation programs do not sufficiently prepare them to deal with testing requirements. Costigan (2002) found that
high-stakes testing policies imposed on new teachers affected their instruction and made them feel relatively powerless. Additionally, the distribution of teachers on a bell curve based on student test scores categorizes educators relative to their peers and fosters a climate of competition rather than collaboration (Scherrer, 2012).

**Impact on learning.** An alleged benefit of high-stakes testing is that it will motivate students to work harder in order to pass the test. Although research does suggest that students may be somewhat motivated by the extrinsic rewards that come from good test scores (DeBard & Kubow, 2002, as cited in Jones, 2007), most teachers have reported that testing has a negative impact on students’ love of learning (Jones, 2007). Indeed, there is a preponderance of evidence that suggests high-stakes testing has negative effects on student learning and motivation. According to Amrein & Berliner (2003), “when rewards and sanctions are attached to performance on tests, students become less intrinsically motivated to learn and less likely to engage in critical thinking” (p. 32). Research also indicates that high-stakes tests may increase dropout rates, especially for students deemed “at risk” (Horn, 2003). When Massachusetts made a passing score on the MCAS exam a graduation requirement beginning in 2002, student graduation rates declined, most significantly for African American and Hispanic students. Of even more serious concern are the reports that high-stakes accountability measures have incentivized schools to keep certain students out, hold students back, or push students into special education so their results are not factored into the school’s reported test scores (Darling-Hammond, 2007). This is a gross social injustice and a testament to the profound shortcomings of high-stakes testing policies.
Limitations. Student growth indicators, the most common of which are known as value-added models, have been used primarily for assessing student growth and teaching effectiveness in elementary and middle schools, and most student growth models only test students in math and reading. Therefore, there is a need for econometricians and statisticians to develop VAMs in subjects and for grades other than those covered by NCLB (Meyer & Dokumaci, 2010). If states or school districts are charged with developing their own measures, such as the common assessments developed by school districts in Massachusetts, generalizability will be limited due to the expected variance in measurements. In addition, it is not yet possible to isolate teacher instruction, especially at the secondary level, from outside factors using quantitative methods. If a student is taking physics and calculus, it is not possible to know whether or not the physics teacher contributed to the student’s knowledge of calculus, or vice versa (Goldhaber, Goldschmidt, & Tseng, 2013). Due to these considerations, this study would be limited to teachers of subjects that rely on standardized test data from MCAS or PAARC assessments to inform evaluation processes.

The rapid utilization of value-added teacher accountability measures is cause for concern since research indicates that different tests yield different results concerning teacher effectiveness. Educators are often not aware of how to use the data to improve instruction, and teachers who receive high scores are not necessarily presenting material in a way that is educationally beneficial for students (Hill et al., 2011). Teachers are privileged or disadvantaged by the students on their class roster, and the regional, cultural, and socioeconomic background of students affects their performance on standardized tests (Amrein-Beardsley, 2009; Darling-Hammond et al., 2012; Koedel &
Betts, 2009; Rothstein, 2009; Steele, Kraemer & Meyer, 2012). Additionally, high-stakes testing impacts curriculum development, teaching, and learning. The curriculum in schools is typically geared toward tested subjects, at the expense of non-tested subjects, teachers feel pressured to prepare students adequately for the tests, and students lose motivation when learning becomes mainly an exercise of skill and drill practice. Finally, VAMs have not been developed in all subjects and for all grade levels, and therefore it is not possible to evaluate all teachers using the same metric. Truly, numerous issues exist with the use of student growth models for teacher evaluation; yet teachers are evaluated according to these methods nonetheless, without regard for how they are affected.

Summary

The bipartisan passage of the No Child Left Behind Act (NCLB) in 2001 established an educational policy trajectory for the twenty-first century that holds states and school districts accountable for tracking and reporting students’ academic progress and emphasizes high quality teachers in every classroom. Race to the Top expanded the role of the federal government by mandating that states in receipt of federal funding overhaul their education evaluation systems to hold teachers more accountable for student learning. While researchers tend to agree that high-quality teaching positively affects student learning, they often disagree over the role of high-stakes accountability policies in teacher evaluation and the extent to which these policies achieve their intended goals (Ballou, Sanders, & Wright, 2004; Darling-Hammond et al., 2012; Hanushek et al., 2005; Hill et al., 2011; McCaffrey et al., 2003).

Indeed, the goals of policymakers in terms of accountability and outcomes are deeply at odds with teachers’ beliefs about the purpose of education (Everson et al.,
2013). Furthermore, teachers are often left out of the conversation about policies that impact their practice (Lee, 2011). Practitioners recognize that student achievement is influenced by the factors such as the socioeconomic background of the students, the condition of the school, the policies of the administration, or the resources available in the classroom (Darling-Hammond et al., 2012; Gamoran, 2001; Kalogrides, Loeb, & Beteille, 2011). Furthermore, these characteristics vary from school to school and district to district, making it difficult to base judgments about teacher quality and effectiveness on student achievement data (Clotfelter et al., 2006). While student test scores may serve a formative purpose for improving instruction, the research suggests they have been used to improperly reward or punish teachers, and the potential variance in outcomes calls into question their reliability as a tool for teacher evaluation (DiCarlo, 2012).

The literature review exposes a gap in the current research when it comes to how changes to evaluation that account for student learning impact the teaching profession. The real value a teacher adds to a student’s learning cannot be measured solely with test scores. As a Race to the Top state, Massachusetts recently revised its education evaluation system to incorporate student learning outcomes in the evaluation of teachers using common assessments (formerly District-Determined Measures) and a Student Growth Percentile (SGP) to determine how much a student achieved relative to his or her peers (Massachusetts Department of Elementary and Secondary Education, 2010). Teachers have had little input regarding these changes, and therefore this study will attempt to bring to light the experiences of public school teachers who have been directly impacted by changes to evaluation requirements as educators of state-tested subject areas. The results will guide district leaders and policymakers to consider the effects of teacher
evaluation changes on the teaching profession. Use of a constructivist approach, embedded in the IPA methodology, will bring rich meaning to those experiences.
Chapter 3: Methodology

Purpose Statement and Research Question

The purpose of this study is to examine how teachers have experienced changes to evaluation procedures that rate educators based on performance standards and student learning. The question for this study is framed to better understand the perceptions and attitudes of middle school teachers of state-tested subjects employed at public schools in Massachusetts as they have navigated changes to evaluation procedures. This qualitative Interpretative Phenomenological Analysis (IPA) study will investigate the following question:

- What are the experiences of educators whose practices have been shaped by changes to the system of teacher evaluation?

Research Design

This study will explore the lived experiences of middle school teachers as they respond to federal, state, and district policies concerning teacher evaluation. In order to investigate teachers’ experiences, it is necessary to understand the meaning it holds for them. Therefore a constructivist social paradigm is required. In adopting a constructivist paradigm, the researcher will attempt to illuminate the realities of the participants as they assign purpose to their experiences (Guba & Lincoln, 1994). The goal of the research is to “search for patterns of meaning” (Butin, 2010, p. 59) in an effort to explain how teachers have responded to changes in teacher evaluation procedures (Ponterotto, 2005). A social constructivist paradigm, also referred to as an interpretivist paradigm, assumes that there is no objective knowledge; rather, knowledge is socially constructed, where the story is defined and redefined according to the diverse perspectives of individuals and
groups. The researcher, already a part of the story about the truth, engages with the participants in an interactive dialogue to co-construct meaning (Butin, 2010; Ponterotto, 2005).

The interpretivist understands that student achievement is impacted by factors that are outside the teacher’s control. These external factors are central to understanding the challenges the teacher encounters in the classroom. Braun (2005) asserts, “it is very difficult for the statistical machinery to disentangle these intrinsic student differences from true differences in teacher effectiveness” (p. 3). The application of a constructivist-interpretivist lens will allow the researcher to examine the various pedagogies that emerge under different circumstances and “thoroughly document the perspective being investigated” (Butin, 2010, p. 60). A constructivist-interpretivist paradigm consists of research that aims to delineate the range of conditions that teachers experience in the classroom, revealing that the same standards may yield vastly different results depending on the subject, class, students, or school. The interviews for this project occurred face-to-face, online, or over the telephone in a setting that was selected by the participants in order to allow them to feel comfortable and confident in their responses during the interviews.

A review of the literature on the reliance of growth models for teacher evaluation reveals that while there are many studies that examine the instruments used for assessing student learning, as well as the impact of high-stakes testing on the curriculum, teaching, and learning, few studies take into account the ways in which teachers experience policies that link student achievement measures with educator evaluation. Furthermore, significant changes to evaluation that have resulted from legislation or mandates have
often been adopted and implemented hastily, with little opportunity for teachers to willingly embrace or adapt to the changes. The researcher, as a scholar-practitioner, is knowledgeable about evaluation processes, both as a teacher and a supervisor, and seeks to give voice to the teachers as they reflect on the ways in which their professional practice has been affected by factors outside of their control.

**Research Tradition**

The study adopts interpretative phenomenological analysis, or IPA, as the guiding research tradition since it aligns well with the researcher’s goal of understanding the experiences of teachers as they reconcile their own perceptions of their teaching abilities with high-stakes indicators imposed on them (Smith, Flowers, & Larkin, 2009). IPA is a relatively new research approach that is “concerned with the detailed examination of personal lived experience, the meaning of experience to participants and how participants make sense of that experience” (Smith, 2011, p. 9). A qualitative methodology, IPA is rooted in phenomenology, hermeneutics, and idiography; therefore, it is necessary to explore the theoretical underpinnings to understand why IPA is appropriate for this study.

The inductive qualitative research method known as phenomenology first gained widespread recognition due to the philosophical writings of Edmund Husserl and Martin Heidegger in the early 20th century. In essence, phenomenology is an approach that focuses on consciousness and experience, and its purpose is to explain the common lived experiences of a group of individuals as they encounter a certain phenomenon. German philosopher Edmund Husserl (1859-1938) is considered the father of modern phenomenology. Husserl, influenced by philosopher and psychologist Franz Brentano’s characterization of *intentionality* “as the fundamental concept for understanding and
classifying conscious acts and experiential mental practices” (Dowling, 2007, p. 132),
sought to understand human consciousness by studying things as they existed in daily life. Husserl’s criticism of the prevailing research methods in human science resulted in the development of phenomenology as an alternative method, still within the positivist paradigm, for conducting research.

Husserl’s characterization of phenomenology connects the researcher with the experience. A phenomenon can only be studied if there is a subject who knows it exists and is able to experience it. Therefore, consciousness and intentionality are integral components in phenomenology. Sadala and Adorno (2002) contend that “there is no consciousness without the world, nor is there a world without consciousness” (p. 283). According to Husserl’s conception of phenomenology, the researcher should rely on reduction techniques, achieved through “bracketing” out of previous understandings, knowledge, and assumptions in order to comprehend the essence of an experience. Central to Husserl’s view is the need to understand lived experiences before these experiences have been corrupted by reflective interpretation (Finlay, 2009; Laverty, 2003).

German philosopher Martin Heidegger (1889-1976) was also a formative thinker in the conception of phenomenology. Like Husserl, Heidegger sought to understand the lived experience. Whereas Husserl concentrated on the description of the experience, however, Heidegger focused on interpretation. This signified a departure from Husserl’s transcendental phenomenological method, which maintains that research is pure description. Heidegger rejected epistemology, the theory of knowledge, in favor of ontology, the nature of “being.” In his work Being in Time, Heidegger advances
hermeneutic phenomenology, which aims to interpret phenomena as they are occurring in everyday experiences. This moves beyond simply describing the experience to making sense of the experience (Dowling, 2007; Reiners, 2012). Merleau-Ponty, influenced by the work of both Husserl and Heidegger, rejected the notion of absolute truth, claiming instead that individual perceptions are influenced by conditions at various standpoints in time. There is no constant; the person and the world are always experiencing transformation and this affects how individuals perceive phenomena (Sadala & Adorno, 2002).

Studying under Heidegger, Hans-Georg Gadamer (1900-2002) developed a dialogical approach “whereby the horizon of the interpreter and the phenomenon being studied are combined together” (Dowling, 2007, p. 134). Gadamer expanded on hermeneutics in his work *Truth and Method*, asserting that an individual’s understanding and preconceived notions are essential to the process of interpretation related to that person’s existence. The ability to examine and comprehend a phenomenon is rooted in the “preconceptions or prejudices…[that] make understanding possible in the first place” (Outhwaite, 1985, p. 25).

The approaches advanced by Husserl, Heidegger, Merleau-Ponty, and Gadamer influence the IPA tradition, which is dedicated to making sense of the first-person experience from the third-person perspective, with a deep appreciation for the confluence of phenomenology and cognition (Larkin, Eatouch, & Osborn, 2011). Idiographic in its commitment to the individual, IPA emphasizes both the participant experience and the phenomenon. Heidegger’s emphasis on the interpretation of a phenomenon and Gadamer’s insistence that preconceived notions are essential to the process of
interpretation are of paramount importance to the study, since an understanding of the experience relies on an awareness that the problem exists in the first place. In seeking to understand how public middle school teachers in Massachusetts experience the new evaluation system, the researcher, who has experienced the new educator evaluation system personally, is committed to situating personal meaning in context through intersubjective inquiry and analysis (Larkin et al., 2011).

**Participants**

The research participants were chosen through purposeful sampling, which is a strategy used in IPA to ensure that those involved have a shared experience (Smith et al., 2009). Additionally, it was essential for participants in the study to have direct experience with the phenomena under investigation, specifically the use of performance ratings and student achievement data in their evaluations. The selected teachers were those who taught sixth, seventh, or eighth grade in the subject for which they were evaluated during the 2015-2016 school year and who were presently teaching the same subject and grade level during the 2016-2017 school year. An exception was made for David, who taught high school science during the 2015-2016 school year and transferred to the middle school at the beginning of the subsequent school year following a negative experience with evaluation in the high school.

Since IPA is characterized by a small sample size, six teachers who fulfilled the required criteria were selected to participate. In an effort to minimize weaknesses that exist with small sample sizes and to gain a variety of perspectives in experiencing the phenomena, the sample intentionally reflects various levels of teaching experience, years in the current position, and curriculum taught. The researcher intentionally selected
participants with at least five years of teaching experience since they have experienced both previous and new evaluation methods.

**Recruitment and Access**

To ensure that this research study adheres to the highest ethical standards in the recruitment and participation of all human subjects, the following process for conducting the research and collecting the data was applied. First, approval for the study was obtained through Northeastern University’s Institutional Review Board (IRB). Following the granting of IRB approval, the timeframe for recruiting participants, conducting interviews, and analyzing the data took several months. Participants were solicited based on Educator Evaluation data available on the Massachusetts Department of Elementary and Secondary Education website for 2012-2013, 2013-2014, 2014-2015, and 2015-2016 school years. In order to recruit participants with a range of experiences, middle school English, math, and science teachers from large and small, urban and suburban, districts were contacted via email using district-published email addresses.

Participants who responded to the initial email were recruited through purposeful sampling. Six participants, two English Language Arts, two mathematics, and two science teachers were selected based on their pre-existing qualifications. The three males and three females who were chosen had teaching experience ranging from six to more than twenty-five years, and they were employed in geographically and demographically diverse districts. Participants were issued unsigned consent forms, notified that participation was voluntary, and told that they could opt out of the study at any time and their identities would be kept private. Interviews were scheduled over several weeks. Participants were interviewed individually in two parts, either in person, via
videoconference, or over the telephone, and they each had an opportunity to review excerpts from their interview transcripts as well as researcher analysis in order to corroborate the researcher’s findings.

**Data Collection**

The study relied on interviews as the basis for the interpretative phenomenological analysis research. Adopting a hermeneutical approach to data collection, which is rooted in Heidegger’s beliefs, the biases and assumptions of the researcher were not bracketed out but instead are considered an important part of the scholar’s ongoing self-reflection and interpretation of the phenomenon being explored. This was done in the writing of ongoing analytic memos that captured this thinking and reflecting. In addition to interviews, the researcher’s personal reflections, a brief biography of each individual that contains information about professional training and teaching experience, and data acquired from observations or documents related to the experience were included (Dowling, 2007; Laverty, 2003; Sadala & Adorno, 2002).

The interview protocol consisted of responsive interviewing, which is a style of qualitative interviewing that emphasizes the relationship between the researcher and subject that is non-confrontational and positive (Rubin & Rubin, 2012). As a crucial part of the interview process, the researcher established rapport with the participants, adhering to semi-structured interviews while allowing the subjects to deviate from the scripted questions. The researcher asked a main question and then continued with follow-up questions to clarify responses and provide sufficient detail for the overarching research questions. Additionally, the researcher relied on probes, such as comments or gestures, to manage the conversation and keep the participant on track (Rubin & Rubin, 2012).
The researcher utilized a journal to reflect on the process and capture meaning throughout the data collection process.

Interviews were scheduled over several weeks, with each participant being interviewed in two parts. Participants were interviewed individually, either over the phone, via Skype, or in person, and they also took part in follow-up collaboration through email in order to corroborate the researcher’s findings. Interviews were recorded using the Rev Voice Recorder: Audio Transcription and Dictation application and a back up recording device. The Rev Voice Recorder app works in conjunction with rev.com, a website that offers transcription services, which was utilized to transcribe the interviews.

Data Storage

A key consideration is the vulnerability of the participants. Therefore, every effort was made to ensure the confidentiality of participants’ responses. Pseudonyms have been used to protect the identities of research participants. Personal notes, journals, audio and video recordings, and transcriptions of interviews were stored in a locked filing cabinet in the researcher’s office, with the key kept in a separate location. Back-up copies were stored using the researcher’s personal Drop box account, which is password protected. All materials will be destroyed within three years of the completion of the study.

Data Analysis

Within the IPA tradition, participant and researcher prejudices toward teacher evaluation inform their understandings of the experience, as does history and tradition with respect to evaluation procedures (Smith et al., 2009). The researcher attempted to describe the participant experiences with the phenomenon by reviewing the data, which
consisted primarily of interview transcripts and the researcher’s notes. The process for analyzing the data followed the steps outlined in Smith et al. (2009).

**Step 1: Reading and rereading.** In an effort to become immersed in the data and recall the settings of the interviews, the researcher listened to the audio recordings of each interview at least twice. The researcher read, reread, and annotated each transcript, using these sources and the researcher’s journals to develop notes reflecting on the meaning embedded in the interview responses.

**Step 2: Initial noting.** The researcher used the notes for crafting a list of thoughts related to the meaning of the participant experience, which then were coded using MAXQDA, qualitative data analysis software. According to the hermeneutic aspect of IPA, the researcher interprets the phenomenon in a way that emphasizes not only the shared experiences of the participants, but also the individual nuances of each participant experience, and the coding of each transcript reflected this methodology (Smith et al., 2009). Following the initial coding, pattern coding occurred, where upon further review of the transcripts, patterns of codes were created that reflect repetitions of ideas into themes.

**Step 3: Developing emergent themes.** The next stage involved looking for connections between emergent themes and grouping them together under a descriptive label. The hermeneutic circle influences this process, as the transcript is reorganized according to the researcher’s interpretation of the data. During this phase, the list of themes was reduced, as some did not complement the emerging structure or were substantiated by weak evidence (Pietkiewicz & Smith, 2014).
Step 4: Searching for connections across emergent themes. Recurring themes comprised a final list, which consisted of subthemes in some cases. Emergent themes, which had been ordered chronologically up to this point, were reorganized into a structure that highlighted the most interesting and important aspects of the participant’s account in support of the research question. The final list of themes was input into a spreadsheet, and quotes from the interviews were inserted as supporting evidence. This enabled the researcher to identify connections between emergent themes to determine their specific functions within the transcript.

Step 5: Moving to the next case. This step involved completing Steps 1-4 with the other interview transcripts. The influence of the previous transcript was inevitable, yet IPA requires each case to be treated independently, and therefore care was given to the subsequent transcripts to allow their individual themes to emerge.

Step 6: Looking for patterns across cases. The last stage of analysis involved writing a narrative that consists of rich, textural descriptions of what the participants experienced with the phenomenon under investigation (Moustakas, 1994; Pietkiewicz & Smith, 2014). This phase illuminated the dual-quality of IPA, exemplifying the ways in which participants “represent unique idiosyncratic instances but also shared higher order qualities” (Smith et al. 2009, p. 101). The results provide insight into the experiences of middle school teachers as they navigate changes to evaluation processes that utilize standards and student learning information as the basis for educator performance ratings.

Ensuring Trustworthiness and Verification

In this research study, the following strategies were applied to ensure reliability. First, the researcher established trust with the participants by allowing the participants to
select the interview method (face-to-face, Skype, or telephone) and location, and engaging conversationally with each participant to collect relevant background information. Second, sections of the interview transcripts were shared with the participants via email to give participants an opportunity to comment or clarify the material discussed during the interview. Additionally, the data was triangulated from multiple interviews with teachers of various subjects and grade levels to develop codes and themes. Triangulation also occurred using data acquired from observations, memos, or documents related to the experience. The researcher employed the validity procedure of member checking, allowing participants to review and comment on the researcher’s themes and categories and incorporating participant feedback into the final narrative (Creswell & Miller, 2000). Participants were given the opportunity to review and comment on the researcher’s interpretations and analysis in order to corroborate the findings or clarify their statements (Creswell, 2013). The double hermeneutic, where the participant and researcher are engaged in a back and forth review of the data, is an important step in the analysis process for ensuring that the researcher’s understandings align with the participants’ understandings of the experience.

In accordance with Heidegger’s philosophical beliefs, the researcher and participants co-construct meaning, engaging in a “hermeneutic circle of understanding” (Laverty, 2003). This interpretive framework allows for the research to evolve over the course of the study, with recognition that the cultural and social contexts of one’s background are key components for understanding. The prior knowledge, biases, and assumptions of the participants and the researcher are inextricably connected to the ways in which they interpret the phenomenon. With respect to bracketing, Gadamer believed
that a knower could not willingly reduce their ability to know something simply “by adopting an attitude,” and attempts to do so were “manifestly absurd” (Laverty, 2003, p. 25). Rather, the researcher and participants interacted to socially construct knowledge about the experience.

The researcher checked for bias by selecting locations outside of the district where the researcher is employed, including participants from a range of disciplines, and encouraging participants to review the data as part of an ongoing process of informing the research study. The researcher’s analytic memos, written during and after data collection, allowed the researcher to periodically flush out ideas about the participants’ experiences. The researcher’s personal notes and memos served as the basis for the analyses, and through member checking, the participants had the chance to review these interpretations to guarantee their accuracy.

In addition, the researcher enlisted the support of colleagues to assist with peer review of the data analysis to challenge the researcher’s assumptions and ensure credibility. Throughout the data collection and analysis process, the researcher regularly met with a small group of colleagues who reviewed and discussed sections of the transcripts, handwritten notes, and the researcher’s interpretations. This stage of peer review allowed for the inclusion of impartial perspectives regarding the research topic and helped to safeguard against researcher bias. As a final phase of the data analysis process, participants had an opportunity to review the researcher’s findings to ensure the interpretations accurately reflect the beliefs and intentions of the participants.
Chapter 4: Research Findings

The purpose of this study was to examine teacher perceptions as they relate to changes in evaluation procedures. The implementation of a revised system of educator evaluation beginning in 2012 occurred with little consideration for how it would impact teachers. The perceptions of educators of state-tested subjects, who may experience additional pressure due to the expectation that their students achieve certain learning benchmarks, are invaluable to understanding how externally imposed mandates affect the teaching profession. This chapter will provide an explanation of the study’s context, an overview of participant backgrounds, including a discussion of their previous and current experiences with evaluation, a discussion of the major themes identified, and a conclusion. The following research question guided the study:

- What are the experiences of educators whose practices have been shaped by changes to the system of teacher evaluation?

Context

Recent changes to the Massachusetts system of educator evaluation compel teachers to demonstrate progress on student learning and professional practice goals, as well as proficiency in the four standards of effective practice, which consist of (a) Curriculum, Planning, and Assessment; (b) Teaching All Students; (c) Family and Community Engagement; and d) Professional Culture. Each of the standards is further broken down into multiple indicators, which are included as part of a comprehensive Teacher Rubric (Appendix A). With the state’s overhaul of the educator evaluation system in 2012, district leaders were charged with implementing the changes and training evaluators and teachers to understand how to use the new system. The new evaluation
tool was first implemented during the 2013-2014 school year, which is when districts began reporting an Educator Evaluator rating to the Department of Elementary and Secondary Education. At the time of this study, the evaluation system had been in place for four years.

Participants were purposively recruited based on their experience as middle school teachers in state-tested subject areas. Six participants—two English language arts, two mathematics, and two science teachers—were chosen to take part in the study. The participants were required to have at least five years of teaching experience so that they could adequately compare the newer evaluation model with previous methods of evaluation. Additionally, participants had to have at least two years of experience in the same role to ensure they could accurately characterize their experiences within their respective schools. An exception was granted for David, who had just completed his first year in the middle school after teaching in the high school in the same district for a number of years. He was chosen due to his vastly different experiences with evaluators in his district.

The participants, three men and three women, represent districts in Eastern Massachusetts that are urban and suburban, and low-, mid-, and high-performing, according to standardized test score reports. They range in experience from six years to more than twenty-five years. This qualitative interpretative phenomenological analysis relied on semi-structured interviews to capture the essence of participant experiences with educator evaluation.

Participants
Evelyn. Evelyn currently teaches seventh grade science in a large urban school district that includes four public middle schools. Her experience includes one year of teaching in a different Massachusetts school district, five years in another state, and two years in a graduate school program, where she also supervised new teachers. She has just completed her second year in her current role.

Previous experience with evaluation. Evelyn characterized evaluation in the previous state where she taught as “very relaxed.” Evaluators would conduct infrequent observations and typically did not hold administrative degrees. In her supervision of practicum students, Evelyn reported that her graduate school held a “healthy view of how observation impacts your practice” and she believes that most students thought her feedback was valuable, which may have been due to the mutual investment between supervisor and student teacher. Evelyn feels that for an evaluation system to be authentic and effective, “the buy in piece is so important.”

Current experience with evaluation. Evelyn recognized that the evaluation tool she used as a supervisor in graduate school is “very similar” to the current tool used by the state of Massachusetts, both of which consist of several key standards and subcategories known as indicators. The way in which the tool is presently being used to evaluate teachers in her school, however, is quite different. As a supervisor of student teachers, Evelyn recalled that the process was “so reflective,” which is in stark contrast to her present experience.

Theresa. Theresa is employed at a middle school in a large, urban school district in the northern part of the state. She has more than twenty-one years of teaching experience, in a variety of roles and school districts. She has taught eighth grade in her
current school for the past six years and presently teaches one history and three English classes.

**Previous experience with evaluation.** As a veteran teacher in multiple roles spanning her career, Theresa has had a range of experiences related to evaluation. She recalled that in the past, she had a principal that told her to “fill out your evaluation and give it me.” Others would complete an initial evaluation, based solely on observation, and then not return again “because they know that you’re doing well.” According to Theresa’s recollections, she typically received “good evaluations” and feels that previous systems of evaluation “used to be more about teaching and helping you.”

**Current experience with evaluation.** In contrast with her experience earlier in her career, evaluation under the current system has “become a time waster,” as well as “data focused.” She believes that “it’s actually very difficult to evaluate a teacher.” Her district experimented with TeachPoint, an evaluation management software program, for a short time to store evaluation documents, but decided to get “rid of it at the end of last year” without telling the teachers in advance, “so it erased all people’s documentation.” This decision confirmed for Theresa how important it is to back up her work using alternative storage methods.

**Matthew.** With six years of experience in education, Matthew has taught in large, urban and mid-sized, suburban school districts. He is currently employed as a 7th grade ELA teacher in a suburban middle school that educates approximately 500 students, and he has been in the role for the past four years.

**Previous experience with evaluation.** In his previous district at an urban school, Matthew recalled a school-specific evaluation tool for the formative and summative
evaluations. As a somewhat new and inexperienced teacher at the time, Matthew believes that he “was lucky in the sense that [his] rating evaluations were proficient.” He contrasted his experience in the former district with his current one, describing the evaluation protocols as “very different” and concluding that he may not be improving as an educator if he had stayed employed there. He also described his previous experience as one that provoked fear due to

an anxiety of the unknown of when they were coming into the classroom, anxious when they were in the classroom, the unknown of what was being viewed in terms of their looks-fors, and then the uncertainty of what the follow-up conversation was going to look like.

Matthew concluded that his experience in his prior district was unsettling due to his status as a novice teacher, which caused him to lack confidence in his abilities, as well as the inconsistency with which classroom observations were conducted.

**Current experience with evaluation.** Matthew’s current district presently relies on TeachPoint to manage documents related to evaluation. He believes that some of the most significant changes to evaluation since he first began teaching are the result of changing vernacular. He credits his skillfulness at adapting to the new evaluation tool to his C.A.G.S. program, in part, where he has been able to effectively “develop a sense of understanding the vernacular and vocabulary” in order to anticipate “the appropriate look-fors when an evaluator’s coming in.” He elaborated, stating

I don’t say that in the sense of when an evaluator walks in, I put on a dog and pony show, but for the sense of when I’m looking at the four standards of evaluation as a professional educator, I’m constantly thinking about what artifacts
speak to the four standards when I’m either planning, or thinking about assessment, or thinking about community engagement, or thinking about professional practice. I think about what are the best artifacts that speak to the standards that then would be able to be highlighted in the evaluation tool. Matthew thinks that the evaluation tool itself, when used appropriately, encourages educators to reflect on what they do in the classroom and facilitates conversations between the evaluator and educator pertaining to elements of teaching. Matthew does acknowledge that a well-trained and competent evaluator is a key factor in its overall success.

Amy. A 7th grade Math teacher, Amy has ten years of teaching experience, two years in her previous district and eight years in her current district, a mid-sized suburban district just north of Boston. During that time, she has primarily taught seventh grade standard and accelerated math, working as a computer teacher for a brief period during her transition from one district to the next.

Previous experience with evaluation. Amy described her experience with evaluation as “simple and straightforward.” Prior to the implementation of the new educator evaluation system several years ago, Amy remembered a previous system that consisted mainly of observations, with a pre-conference that included a discussion of the lesson, an observation, and a written report that required acknowledgement from her. Of the entire process, she recalled, “I don’t remember it being a lot of work.” She characterized previous evaluation procedures as less formal and more focused on classroom observation.
**Current experience with evaluation.** In contrast to the observation-only method of evaluation, the focus of the current system, according to Amy, “is more spread out.” She referred to the more comprehensive evaluation tool as “improved” since it takes into consideration other elements of teaching that extend beyond the classroom: “You have your observations… professionalism has its own category. Family communication and all those other aspects they are bringing into it.” She admitted to spending “an above average amount of time putting my, quote unquote, evidence together because it is very important to me to show what I do. I always go beyond the minimum required… [because] I’m very aware that they want to see more now.”

Amy described her training in the new evaluation tool as “very clear.” A professional development workshop, led by her principal, communicated the district’s expectations for evaluation and emphasized that administrators “do want to see evidence for various categories.” Teachers subsequently were trained in TeachPoint, which Amy finds “pretty easy” to use, allowing her to “upload things instead of creating some sort of binder.” She views the current system as “more relaxed” even though she is “doing more work” now because she feels “more comfortable and relaxed about what I’m asked.”

**David.** A veteran Science teacher, David has more than twenty years of classroom experience in K-12, with eighteen of those years in Massachusetts. At the time of the interview, David had just completed his first year in his current role, which is eighth grade Science, having transferred from the high school in his district, a homogenous suburban district, where he had taught for many years. He was selected to participate based on his evaluation experience, which contributed to his decision to transfer.
Previous experience with evaluation. David reported that he did not have much of an evaluation mechanism until the last three or four years. Prior to the implementation of the recent educator evaluation system, David recalled having his “boss come in and say, look, I’m short on time, would you please write your evaluation for me?” Previously, he described the evaluation as consisting of “mostly a write up based on observation.” He also reported that he has “never had a science person evaluate” him, and he suspects that the unfamiliar nature of his subject kept many supervisors from observing more than once in a year.

Current experience with evaluation. David’s district uses TeachPoint to manage evaluation documents. He is evaluated on all thirty-two indicators included in the rubric, as well as on his progress toward meeting his student-learning goal and professional practice goal. His goals are reviewed by his evaluator during the goal setting process and approved if they are deemed “practical.” David discussed the district’s reliance on a software program called MasteryConnect to track student learning growth data.

Jack. Jack is a 6th grade Math teacher with more than twenty-five years of experience teaching at the elementary and middle school levels. He is employed in a large, urban district on the South Coast. Jack is worried about proposals to use student test scores as a condition of employment, which factored into his decision to take part in the study.

Previous experience with evaluation. Prior to the roll out of the new evaluation tool, Jack recalled being evaluated based primarily through observation, where “the principal would come in, or whoever the evaluator, would stay for a few minutes, and they would talk to us after” to debrief what they saw. He described the involvement from
the educator as minimal, consisting of little more than a pre-conference, post-conference, and evaluation acknowledgement. He recalled the use of portfolios as well, but noted that the observation component was fairly consistent.

**Current experience with evaluation.** Although the observations have not changed much under the new system, Jack reported that it feels “more threatening” than in the past. During the past year, his evaluator, the principal, came in “once announced, twice unannounced” and stayed for the entire class period. Jack typically receives “little notes” about the visit in his mailbox as feedback. Jack explained how the evaluation system as whole feels like “more work and more pressure on the teachers.”

**Data Analysis**

An analysis of the interview transcripts led to the identification of one hundred seventeen initial codes, which were reduced to seventy-two, and then further reduced to forty-two, after similar codes were combined and the least repetitive, those appearing fewer than five times across the transcripts, were eliminated. The following themes capture the essence of the participants’ experiences with educator evaluation:

- Accountability driven
- Structured
- Time consuming
- Inconsistent
- Relationship dependent
- Political

The remainder of the chapter will address participants’ experiences related to these themes.
Accountability driven. Participants readily related top-down accountability measures, the product of government regulation and policy, to new evaluation methods that place educators under greater scrutiny. While participants could capably articulate their beliefs about the purpose of evaluation, the goals they identified do not necessarily align with the reality of their experiences. Matthew emphasized the importance of calibration among evaluators and conversations between the evaluator and teacher to push self-reflection. Evelyn also discussed the importance of reflection, identifying with her graduate school’s philosophy regarding “how observation impacts your practice.”

David stated, “I think the purpose of evaluation is to help somebody grow professionally…[and] the evaluator is mostly a person that can guide a teacher into becoming a better teacher, a more effective teacher.” According to Theresa, evaluation should simply be about “teaching and helping you.” On the other hand, Amy expressed skepticism that the current system accomplishes any of these things: “I don’t think the reality of the system improves anything at all.” Participants seemed in agreement, however, that the Department of Elementary and Secondary Education’s evaluation requirements supersede any beliefs school leaders or teachers have about evaluation’s purposes.

As a condition of Race to the Top, Massachusetts had to develop an evaluation system that made teachers more accountable for student learning, and participants have felt the impact of this change. When sharing their experiences about the relationship between evaluation and test scores, those participants who teach in districts with a greater percentage of high needs students or with average or below average standardized test scores emphasized how test score reports influence school policies and steer evaluation
procedures in their schools. Theresa thinks that evaluation conveniently gives school districts an option to rationalize low test scores by equating student performance with teacher quality. Although she feels the real reason for low scores has more to do with “higher free lunch, higher this, and not support,” the state’s leveling of schools as high-, mid-, or low-performing based on student test scores creates burdens for district leaders, who are then compelled to develop solutions for closing the achievement gap. According to David, “sooner or later administrators are going to have to justify why a certain segment of the population is not getting it” and ultimately, “it comes down to what the MCAS score is.” Theresa effectively highlighted the connection between test scores and evaluation:

how can you be a level three and have all your teachers be proficient? The principals felt the need to prove that teachers weren’t proficient… If you weren’t a level one school, they felt like they had to have teachers that were failing to show why they had a lower level for their school.

Theresa observed that the hold harmless position of the state due to the PARCC trial in 2015 and 2016 and transition to Next Generation MCAS, which meant the district would be held harmless for any negative changes in school and district accountability levels through the 2017 testing cycle, reduced the burden placed on districts to raise test scores at all costs, which was also felt by teachers. While Theresa’s school is rated as level two, the high school is labeled a level three, so in the past, the district as a whole was rated a level three. In the hold harmless phase, her school, as a level two school, is less impacted by the high school’s lower rating. She determined, “As long as we’re in this no harm, no foul, I think the pressure’s off the principals, or at least in my district
anyway, to be as careful or as concise about the evaluations. Whereas when the district was level three, even though we were level two, they felt more pressure.” As a teacher in a district that also trialed PARCC, Matthew discussed being “in limbo the last two years because we haven’t had reports coming back to use at the middle school level because…the PARCC data wasn’t reported back.”

Despite the hold harmless phase, some participants still reported feeling pressured by district and school leaders to raise achievement. Evelyn relayed the feeling that her principal “has to justify his existence with test scores, and…it feels like that’s all that’s important.” She explained how her teaching assignment was altered to include a science intervention class at the beginning of the school year in an attempt to boost the school’s science MCAS scores, which were “atrocious.” She recalled how her principal “took the top group, only the top group because he thought they had the best chance at learning and improving…because he thought he could get more kids into advanced if I worked with them all year.” She further discussed how other areas, like behavioral or student social emotional issues, are often overlooked or diminished because the school’s focus is “entirely based on testing.”

David also commented on how state directives impact district policies, noting, “Any time you go from top-down, you get exactly what you get in a top-down system; you get a bureaucracy.” Furthermore, you fail to take into account “what the perspective is in the classroom, from the classroom teacher, and what they’re seeing.” The result has been an evaluation system that places pressure on district leaders to reduce the achievement gap, as measured by performance on state standardized tests, who then pressure building leaders to raise scores, who, in turn, pressure teachers to be more
accountable for student learning gains. Evelyn also characterized it as “a top-down system” that “trickles down to us.” She acknowledged that her principal is just pushin’ it on down to us. Like, “Make it happen for me, guys,” because he’s got someone else ahead of him, someone that has a superintendent pushing pressure on all of them, but like if the kids don’t learn astronomy because this teacher can’t teach astronomy, the kids will not get any astronomy questions right on the MCAS.

Jack also reported feeling testing pressure in his district, which includes a significant percentage of high needs students. Jack explained, “No matter how much you work with them, they may not succeed as well as they should in the MCAS.” He elaborated, describing his students from the current school year as “low, low level,” finding it “unfair” to “attach a score.” Like Theresa, Jack suspects that “the state is coming down hard” on his district, “making them say that no one is perfect just because the scores are not good.” While Amy acknowledged the use of MCAS data to inform her goal writing, she did not convey feeling the same degree of pressure to raise student achievement as participants in schools with more high-risk student populations. She admitted to reviewing MCAS data “just out of curiosity” for her own benefit, without any real understanding of how administrators in her district use testing information. Yet she also expressed concern with the present system since “people’s jobs are on the line because of it.”

Significantly, due to accountability driven policies and educator evaluation changes, participants have experienced a shift in recent years that places more responsibility for student performance on the individual teacher and less on the student.
Theresa reflected, “When I first started teaching, seventy-five percent of the work was the students, and twenty-five percent were the teachers. Now I say it’s more like ninety percent are the teachers, and ten percent is the students. It’s a big shift.” Jack feels that students are “not taking ownership of their learning. I think once they take ownership of their learning, then I think it’d be a little easier, but they don’t take ownership of it.” David agrees, viewing evaluation as an instrument that can be used to “say this educator’s inefficient, he didn’t do this. It seems to shift some of the responsibility away from where it really needs to squarely land upon, which is the student.”

**Structured.** While on the one hand, participants characterize the evaluation model as top-down and bureaucratic, they also recognize its potential. Their experiences signify that it offers an underpinning to the entire process of supervision and evaluation by defining the major components of effective teaching, which did not necessarily happen previously using the observation-only method of evaluation. David described the current system of evaluation as “comprehensive” since “it really does outline what the heck you should be doing as a teacher.” In her description of educator evaluation in Massachusetts, Evelyn acknowledged that the use of the rubric to evaluate teachers on their progress in meeting certain standards is “very structured” and “transparent.” Amy assumes “very little” about her teaching gets lost in evaluation since she’s “uploaded so much” into TeachPoint. She described the ratings as “accurate” due to her ability to control how many artifacts she submits as evidence of “parent engagement, parent communication,” lesson plans, and records of “different things that are going on in the classroom.” Matthew views the tool itself somewhat positively, since it compels his evaluator to frequently visit his classroom to get a pulse on what’s happening and coach.
him to reflect and improve. Additionally, he reported that it is evident that “evaluators have been calibrated using the tool and go on instructional rounds together and then debrief the experiences together” since the implementation of the tool in his current district appears to be fairly consistent. Matthew admitted, “I feel that I’m developing more professionally from the current model… I feel like there’s more work, but I’m getting stronger output and more gains professionally because it’s working appropriately. I know that the tool’s working for me.”

**Goal setting is required.** In some districts, educators exercise considerable autonomy over the goal-setting process. Those who developed individual goals reported greater satisfaction with the process. Evelyn mentioned that she had a choice in the development of her student-learning goal, which she elected to base on a science practice skills rubric. She noted, however, that the principal altered her teaching assignment at the beginning of the year to include an intervention class aimed at raising test scores, and had she known this prior to goal setting, she may have based her student learning goal on assessments for those students or eighth grade MCAS scores. Jack also developed an individual goal, describing the goal-setting process as “always challenging,” yet cooperative, as “the principal did help… with writing the goals.” Amy identified MCAS score reports, distributed by her building principal, as “the biggest piece of data” used to inform the creation of her student learning goals.

Some districts have tried to streamline evaluation procedures by encouraging the development of team goals, common goals established by two or more teachers. Those participants who develop team goals communicated higher levels of frustration than those who establish individual goals for several reasons. Theresa admitted that the
development of team goals is “complicated” since teachers on a two-year evaluation cycle are assigned to either year one or year two. Those on year one must establish new goals, while those on year two are still working on goals created the previous year. Additionally, educators who teach common curriculum may not be assigned the same evaluator, which may result in different expectations. Being at different points on the evaluation cycle, combined with varying expectations from evaluators, hinders collaborative efforts among teachers when it comes to evaluation.

Matthew conveyed feeling discouraged by his superintendent’s decision for teachers to develop team student learning goals since the district views the grade seven test score data holistically instead of breaking it down by individual teacher. Matthew finds it unfair that for the past two years, he has “carried the weight of all of grade seven.” If this is a common practice across districts, Matthew believes that “bad teachers are still going to be able to keep their jobs” since it will seem as though their test scores are satisfactory and they will appear to be meeting their student learning goals. Matthew’s individual analysis of his own students’ test performance has made him aware of this inaccuracy.

_Student data informs practice._ Some participants acknowledged that aspects of the educator evaluation system positively impact their teaching. Aware of the need to measure student-learning growth, Theresa developed goals to determine reading gains in her students. The district had invested in a literacy curriculum that Theresa felt did not produce enough data, so she went online and found a reading test to establish a baseline for her students early in the year. Upon administering the test at the end of the year, she discovered that many of her students had “increased their reading levels,” which helped
validate her efforts in the classroom. David also commented on his district’s use of student learning data, discussing the various methods his school is using to assist teachers with creating common assessments to measure student growth. He admitted that programs such as MasteryConnect are “valuable,” but “overwhelming” and “time consuming.” He questioned whether he is “the type of teacher that can effectively use” student growth data from MasteryConnect to increase students’ understanding of a myriad of concepts. Amy also explained that her student-learning goal was based on a pre-test/post-test where she was aiming to have seventy percent of her students earn a seventy percent or higher on the post-test. She reported that she’s “met every goal,” but she’s been reluctant to create goals that are too challenging because she worries how she might compare to teachers who establish “easy” goals. Amy also confessed to “self-reporting” on her student growth data, stating, “No one collects the tests. I have them in my closets, but no one’s asked for them.”

**Time consuming.** While acknowledging that the evaluation system overall is more structured, participants also reported feeling consumed by the vast amount of documentation required. Theresa described the process as a “time waster…me collecting all this data and collecting all this information and spending all this time for my evaluation, which should be simple.” Matthew characterized the current model as “overwhelming in the sense of documentation and procedural stuff… [such as] the goal settings, the deadlines, the reminders, the constant signing off of materials.” Amy admitted to spending considerable time gathering evidence. She recognizes that not everyone can spend the time equivalent uploading artifacts due to other obligations or because to others “it’s not as important.” As a result, she feels “it’s not a true reflection
of what teachers are doing.” David believes that the evaluation system is “too big” and “not practical.” He is convinced that we have “an evaluation system that’s damn near impossible for an evaluator to do correctly” and because of that, “there’s a lot of rubber stamping.” Evelyn stated, “I feel like I’m jumping through hoops.” Amy concurred, inferring that the evaluation system does not actually improve teaching “because there’s so much focus on checking the boxes, and even on the admin side… there’s so much focus on just showing they’re doing their job,” whether or not it is actually helping teachers.

Several participants discussed being required to submit evidence for every single indicator. David explained,

the rubric is laid out, it’s a rather long rubric, there’s thirty-two different items. Basically, I believe that you’re assessed on all thirty-two areas and you have to show evidence of having to meet expectations in all thirty-two of those areas, whether they fit your goal or not.

Theresa described spending as many as forty hours on a binder of evidence in past years when her district required evidence for each indicator. She credits technology with simplifying the process of evidence collection somewhat since she can upload documents into Google Drive as part of an ongoing process throughout the school year. Some participants suggested that technology has not led to improved communication or feedback from evaluators. Theresa explained, “The only time we get feedback is if we get a negative parent phone call.” She was quick to rationalize her administrators’ behavior, however, speculating that the lack of feedback has more to do with the demands of the job as opposed to an intentional rebuff. She stated, “In their defense, I
think it’s because they are probably so inundated… They’re probably so inundated with the random parent requests. I would not know how they could possibly do all of it.”

David also believes that administrators cannot handle the evaluation demands: “You can’t take thirty-two areas and do observations on half the teachers in the school, because the two principals divide it up, their job’s too overwhelming. It’s not practical. You’ve got to make something workable.”

Amy’s experience corroborates the theory that administrators are overwhelmed by the requests placed on them due to both the compulsory evaluation requirements imposed by the state and as school leaders. Amy perceived evaluation to be something administrators suddenly become concerned about when they have a deadline. She recalled,

No one this particular year has been in my room until, I want to say, the week the evaluator observation was due. And that day, he came into my room and observed me and he just told me two days before… That same day, the math department head was in our school because she was going to be doing the exact same thing and waited until the last day, literally.

Amy’s opinion of evaluators in her school is that they are “doing the minimum. And they’re waiting until the last second and you can tell. That’s my experience.”

Participants communicated that districts, in recognizing the complexity of evaluation requirements and the increased demands on teachers and evaluators, may attempt to ease the workload by reducing the number of overall indicators for which teachers must submit evidence and utilizing evaluation management software. Evelyn detailed how her district identifies specific “power standards” for teachers to focus on,
which allows her to more readily target certain areas. Amy discussed how her district also highlights “eight specific indicators,” and she made sure that she submitted evidence for each focus indicator and then “added more stuff in.” Theresa reported that “this year was a little better because they said we only had to do the third and fourth for proof, and then the student and professional goals, whereas previously you had to do every single item.” Jack’s district also identifies specific indicators for teachers to address, yet this has not necessarily made evidence gathering less of a burden. He described the process for submitting evidence: “I have to scan it, put it in my hard drive, in my Google Docs, make sure it’s all there. It’s a long process. It’s a pain.”

Districts rely on evaluation management software as another method for streamlining evaluation procedures and storing documents. Amy, David, Evelyn, and Matthew all discussed using TeachPoint for housing evaluation materials, while Theresa described using TeachPoint for a short time before her district abandoned it, probably because “it costs money.” Theresa and Jack also discussed using Google Docs rather than a formal management program. Although Matthew’s district uses TeachPoint, he explained that his evaluator also created a collaborative Google Docs folder to make the process of evidence submission less threatening. While Theresa indicated that Google Docs works for her in terms of uploading evidence, she believes that “TeachPoint is actually more secure” since “Google Docs is more of a fluid document.”

Participant experiences with TeachPoint are mixed. Amy described TeachPoint as “pretty easy” to use and feels “grateful that I can just upload things instead of creating some sort of binder. I love the technology piece. I love how easy it is.” Evelyn called TeachPoint “very structured” and admitted to feeling satisfied with being able to “go in at
any time and see what they put and…make edits.” Yet she also communicated feeling uncertain when using TeachPoint due to a lack of follow up from her evaluator, explaining, “Only for the formal announced observations do I get a follow up.” Formal guidelines about how much evidence should be submitted are also lacking in her district, so to compensate, Evelyn will “upload some documents and write about them, but it takes a long time. I’d rather be doing my job than uploading evidence.”

David described TeachPoint as impersonal, but attributes it to the present state of evaluation in general. When “you give these principals so many people to watch and evaluate in so little time, along with all the other things they have to do within the building…they can’t do justice to the job they’re supposed to be doing.” In comparison to evaluations that “used to be conversational,” the increased requirements have caused districts to adopt programs like TeachPoint to help manage evaluations, and as a result, evaluators “walk in with an iPad,” observe for fifteen minutes, post their observation online, which generates an email that notifies the teacher, who then reads and comments on the observation. In many cases, the entire process feels rushed because the evaluators are always “in such a hurry” to get it done.

**Inconsistent.** For teachers who have been employed in multiple schools or districts, inconsistencies in evaluation protocols are glaring. For those whose employment has been fairly constant, the discrepancies may be subtler; however, participants are in agreement that evaluation procedures differ depending on evaluator assignment and place of employment.

**Differences among districts.** Having experienced evaluation in an urban and suburban district, Matthew thinks that if evaluators are not having instructional practice
conversations, “the evaluation rating is dependent… on the outlook and perspective of the evaluator coming into your classroom… and the relationship that you previously have built and established with the evaluator, which takes away from the purpose of an authentic evaluation.” Matthew expressed frustration with the use of the ratings system, which he believes is “flawed.” He is disappointed that his district “has decided to not let an overall rating or individual standards to be a rating of exemplary.” Theresa reported that in her district, teachers were told they “were never going to be able to get exemplary,” so she was “shocked when we actually did get it” for parent communication. Teachers’ understandings are that “you almost have to get published” to earn the exemplary rating, and so Theresa figures it makes sense to “just focus on proficient because they’re not going to give me exemplary no matter what I do.” Matthew believes that limiting teachers by communicating that the majority are proficient, especially when the range is so wide, does a disservice to the profession by unintentionally reinforcing mediocrity. He thinks that his superintendent will not modify his position on the ratings until the state is able to highlight, define, and showcase what the word modeled means. That’s a little bit irritating, because as a student, I strived for the “A.” When I go to work each day, I strive for the “A”… and I know damn well that I’ve worked harder than three members of my department, they’re also seeing the same ratings as me, because needs improvement to proficient is very different. They won’t go on an improvement plan, because they’re enough to be proficient. Yet then break down what proficient is. It’s annoying.
Matthew feels that the ratings are not consistently assigned across districts. He struggles with the notion that “an urban teacher can see exemplary in all four standards, and I work my tail off and probably work harder than I did in the urban district and have never been able to see more than proficient in individual standards or overall rating.” He feels that “true rating” is often lost in the current system, since teachers who are less effective may still earn a proficient rating due to the extensive amount of time and energy required to place a teacher on an improvement plan. Matthew concludes it is easier to “call it proficient” than “develop an improvement plan… [because] it is so much more work than just a normal evaluation.”

Jack also commented on the evaluation ratings, voicing concern with his district’s decision to label all non-professional teachers, those with fewer than three years of experience, as “needs improvement.” He stated, “I don’t think that’s fair. I’m sorry, because there are first year teachers that are remarkable, and they’re just getting needs improvement. To me, it’s a let down for them.” Jack is uncertain why novice teachers automatically receive a needs improvement rating, characterizing communication about evaluation in his district as “lacking.” He suspects the ratings are due to external pressure from the state “because the scores are not good.” He assumes in districts where “the test scores are probably high, they’re probably not, they probably don’t stress over it, but if you’re in an underperforming district, then they’re more apt to be a little… tougher.” He indicated that these types of pressures negatively affect the morale of the teaching staff, leading to “no stability” and consistent turnover. While Matthew agrees that the ratings are unfair, his experience suggests that teachers in high performing districts are also concerned about the ratings. Evelyn commented that she has heard “people stress about”
the ratings, comparing themselves to other teachers and discovering, “I only got proficient, that person got exemplary.” Amy admitted that she is curious about who views her evaluation materials and wonders how she compares with other teachers. She confessed she would be interested to know, “What are they doing with all of this? Who’s seeing all of this? How many teachers are getting put here, here, and here? And what is the breakdown?”

**Differences within districts.** Even within the same district, evaluation procedures are not consistent. Participants identified poor communication as a significant reason for the incongruity. Evelyn reported frustration with the lack of clarity about teacher expectations. Although quite adept at uploading evidence into TeachPoint, when it came time to submit evidence for her Formative Evaluation in January, there was “absolutely no guideline,” and Evelyn struggled to answer questions like, “How much do they want? What do they want?” Jack expressed uneasiness with the reality that “they’re always changing everything,” describing a lack of awareness concerning his district’s adoption of certain policies, such as the automatic non-professional teacher rating of needs improvement. Matthew also conveyed feeling confused about his district’s policies concerning ratings since he “was exemplary in professional school culture and family community” but now he’s considered “proficient in both of them.” His current evaluator, however, was told not to rate anyone as exemplary, so no matter how hard he works, he will always be considered proficient. Yet even his evaluator confessed that she is not aware of the reason for this decision, which was made “before her time.”

Evelyn feels that insufficient communication between building administrators impacts evaluation procedures in her school. The principal and assistant principal are
each responsible for evaluating half the teachers; the principal’s feedback is brief, only “a couple sentences,” whereas the assistant principal completes assessments that are “pages long.” The varying nature of evaluations in Evelyn’s school does not go unnoticed by the teachers. Evelyn recalled her coworkers’ comments when they learned who would be evaluating her: “Oh, you got her. She’s tough. She’s gonna ride you… oh, if you got the principal, it’s going to be much easier.” Yet she explained that some teachers are “craving more feedback,” but only getting the minimum. Theresa explained that evaluation responsibilities in her school are divided among the principal, two assistant principals, a dean of students, and a special education administrator. These evaluators have differing expectations when it comes to what they want from teachers. Theresa stated, “my principal likes long, written explanations for things, whereas the upper school assistant for us says, I want the bare minimum… so it depends on who you have.”

Evelyn also spoke about a communication deficiency related to evaluation expectations. When she was in her first year at her current school, she recalled asking more experienced teachers questions about how to upload evidence and how much she should upload, only to learn that teachers are doing entirely different things in terms of evidence submission. While some teachers do use TeachPoint, others collect documents in a binder “because they don’t know how to upload to TeachPoint” since the district has not invested time in training teachers in its use. Since she “cannot get a straight answer about how much evidence” to submit or what she should be covering, Evelyn stated, “I just try and use some common sense, and I ask questions and hope I got it.” These accounts from participants reveal a lack of consistency in the implementation of the evaluation tool across classrooms.
**Relationship dependent.** Participants identified the evaluator relationship as a crucial component of evaluation. The degree to which participants willingly collaborate and depend on coworkers for support appears related to their overall experience with evaluation in their schools.

**The evaluator/teacher relationship.** Each of the participants spoke about the significant role the evaluator plays in setting the tone for evaluation. Jack described his relationship with his evaluator as “a good relationship” because he always does what he is told. Matthew characterized his evaluator as a “coach” and “professional,” whose perspective is “necessary” and “helpful.” Amy indicated that she believes the relationship between evaluator and teachers is “a critical component” to the evaluation process. David is convinced that his positive experience with his current evaluator has “more to do with the type of person he is” than anything else.

Matthew praised his evaluator’s efforts to make evaluation an authentic process, explaining how she observed him six times over the course of the year when she was only required to visit his classroom once. Matthew detailed a typical observation:

No longer than fifteen minutes. Not shorter than probably, like, seven. Never came in with an iPad. Always came in with a notebook. Would do things like either sit in my room and engage in conversations with the students, collect samples of what I was doing, look at taking pictures of my agenda or artifacts in the room. They all cycled back in a positive conversation and showed up somewhere in the tool. Direct quotes from students, where they were able to verbalize the objective of the lesson or pictures of the agenda that highlighted
planning and curriculum mapping. It’s just another pair of eyes to think about the big picture, what works best for students.

After each observation, Matthew and his evaluator conference and she prompts him to consider what he thinks went well with the lesson and what he would want to change if he taught it again. These “honest, constructive conversations” push him to reflect on his practice and foster a “collaborative” relationship built on “trust and rapport.” Following the conference, Matthew’s evaluator will submit a write up of their conversation in TeachPoint, and he will have an opportunity to comment on her summary. He credits his evaluator with pushing him to improve as a teacher through ongoing reflection.

David also described a constructive experience with his current evaluator, who he referred to as “very fair-minded.” He elaborated, stating,

He’s very good. He’s very fair. He takes it very seriously. He’s a very fair, earnest individual, but he’s very open, too, and if I have something to say, he’s going to listen to my opinion, and my opinion actually counts for something. So, I value that very strongly in somebody.

David has known his evaluator “for a very long time” having previously taught together at the high school, yet David is convinced that his supervisor’s personality characteristics are the most important factor affecting his leadership.

Amy reflected on her relationship with her previous evaluator, a former math teacher, describing it as “very close.” Amy felt the previous evaluator “had a good pulse” on her classroom due to her content expertise, which resulted in the writing of “a nice narrative” about what she had observed. Amy stated, “When she came into my classroom, I knew that she got it.” Evelyn recalled a similar experience in a previous
district where her evaluator, as the science department head, provided content specific guidance: “They were the most helpful. That’s the best observations I had. Someone who is a content expert with teaching experience.”

Theresa credits her ability to effectively navigate evaluation procedures, in part, with getting to know her principals. According to Theresa,

My current principal, the one evaluating me, she wants to be in the paper in a positive way and she wants the community to see her in a positive way. I just make sure that when she asks us to do an essay contest, they do the essay contest, and I have a winner so she’s in the paper; or we had the parents come in and we took pictures and she was in the paper. I found that for me, my evaluation is proficient because I know her game.

Her knowledge of what her principal is looking for has influenced some of her decisions beyond curriculum and instruction. Theresa stated, “last year we were told that if we didn’t include four parent contacts during the school year, they wouldn’t sign our evaluation… so we put that in.”

Evelyn discussed the need to be mindful when making requests of either her principal or assistant principal since their relationship is often combative, and they are “never in communication with each other.” She described her relationship with her evaluator, the assistant principal, as “developing” since she seems to have “very little experience” and this creates the potential for conflict with teachers who do have experience. Matthew also communicated this sentiment about his evaluator, having recognized that more veteran teachers in his department “don’t respond well to her.” He suspects this might partly be due to the “huge drama in the department” that occurred
when his evaluator, as an external candidate, was hired as the English department head over two veteran high school teachers who were already employed in the district and had been encouraged to apply for the position. This created tension in the department and makes it more difficult for his evaluator to get “buy-in from those ten colleagues” in the high school. Ultimately, participant responses indicate that the personal relationship between the evaluator and the educator does have an impact on the evaluation outcome, and a more positive and trusting relationship translates into a more gratifying experience for the teacher.

**Relationships with colleagues.** Several participants indicated that strong relationships with colleagues are beneficial, especially in school climates where evaluation has been less positive. Theresa explained that when her building principal insists that teachers engage in certain activities, like including more parent contacts, her colleagues adopt the attitude that “we’re all in this together!” This type of thinking fosters a more productive and collaborative work environment, where teachers work to accomplish an initiative where they all “get credit together.” Jack also noted that he “collaborate[s] all the time” with his grade six colleagues and that “for the most part, the teachers do support one another.” These relationships generate stability for educators in an environment that can feel “threatening” and “always changing.”

Participants who recounted more positive experiences with evaluation rely less on colleagues for support. Matthew reported feeling discouraged at the prospect of receiving the same rating as his coworkers who “aren’t striving and working hard.” As a teacher who consistently goes above and beyond what is required, Matthew admits this can create conflict between teachers, with some colleagues “calling other colleagues try-
hards,” which negatively impacts morale and encourages competitiveness. Amy views herself similarly, describing the other seventh grade teacher in her school as someone who “wants to do what I’m doing because it’s easier to have someone else [do the work]… but I don’t think it’s because he really cares to collaborate because he doesn’t.” Evelyn described a longing for teachers to collaborate with, explaining how in middle school, “you are so isolated.” As the lone seventh grade science teacher in her school, Evelyn discussed having “no one in my content area to talk things over with. No one understands what I’m teaching, and they’re in different schools.” She explained that the once a month meetings with science teachers across the district are not enough to develop a productive work relationship. She equates collaboration with personal growth as an educator and feels that she is “not gonna grow as a science teacher in [her] current position.” While she may not have anyone in her school with whom to naturally collaborate on curriculum, Evelyn did report that she was placed in the eighth grade science teacher’s classroom “this year for two periods a day” because his teaching is “bizarre” and they hoped she could serve as an informal mentor so that “he would improve,” and therefore the MCAS scores would improve. According to Evelyn, this attempt at forced collaboration was not effective since “he doesn’t want me to mentor him.” She did not know the outcome of his evaluation, but she did believe that the school should “just hire a new science teacher” because “he’s terrible.”

**Political.** Participants shared a common belief that the educator rubric conveniently empowers administrators to retain the teachers they like, dismiss those they do not, and manipulate teachers into doing things they may ordinarily resist. Although
none of the participants think that evaluation should be subjective or coercive, it is clear from their experiences that it has the potential to be both these things.

Participants spoke about the need for an authentic system of evaluation that recognizes quality teaching and develops teachers that need to improve. In some cases, evaluation may be used to expunge poor-performing teachers from a district. David admitted, “I don’t have a problem with that. They need a mechanism to push bad teachers out.” More often it seems, however, that participants view evaluation being utilized as a tool to embolden administrators and subdue more outspoken teachers. Theresa revealed that she cannot escape the feeling that the current system is more “about catching people making mistakes.” Whereas previously an administrator and a teacher may have disagreed or had a difference of opinion and administrators “didn’t know what to do, or they didn’t have a plan,” now they can use the rubric and the threat of a performance plan to silence “people who aren’t status quo.” Evelyn concurred, noticing that teachers who are more vocal members of the union received ratings of “needs improvement on stuff,” even if their students earn high test scores. According to Jack, “Sometimes I feel that they want to find fault with what you’re doing in the classroom and it just honestly doesn’t sit well with me, but that’s the truth.” Theresa explained that teachers have become reluctant to speak up since

there’s been so much retribution with it; it was almost like they had a set system they could go after somebody with, and they did. I watched people go through it. Being… I don’t want to say picked on, but picked apart would be a good way to put it, picked apart. Some of them just because they didn’t like what the administration was saying.
Theresa recalled how a colleague “replied all” to an email from the principal about a scheduling issue, and “the principal saw that as a personal attack,” so the English teacher ended up on a ninety-day plan and “was almost out the door.” In another instance, Theresa witnessed a colleague negatively evaluated for arriving to school several minutes late, yet explained, “He’s also the person who stays until five-thirty or six at night with students.” Ironically, it was the teacher’s high test scores that saved him when he was called in to meet with his evaluator and district leadership. The evaluator was hyper focused on the teacher’s punctuality, even though “his test scores for science have been the highest in the district since the MCAS have happened, all the time, every time, and we have the least able students, so it’s pretty incredible.” In both cases, the evaluator attempted to use the evaluation punitively to admonish teachers for minor infractions, but their English and science MCAS scores ultimately saved them both.

A common feeling among participants is that evaluators often base decisions on preconceptions. His own experience leads David to conclude that evaluation is “political in some schools” but he acknowledges, “It’s like any other system. You set something up to work in a specific way, and you trust the people in charge to actually run it that way. So, sometimes you get that, and sometimes you don’t.” David’s personal experience with his prior evaluator confirms that some supervisors appear to use the rubric ratings more as a political tool than an authentic assessment of educator effectiveness. David confided that a negative experience with his high school evaluator contributed to his decision to transfer to the middle school: “I felt I was very unfairly evaluated last year. There’s a major disagreement between me and the person that was doing my evaluation, so I didn’t see that as being anything that was reconcilable, so I
decided to move on.” He reflected on the experience further, explaining that his evaluator, the assistant principal, “would come in one time for an evaluation per year… [for] maybe five minutes. There was no pre-evaluation meeting. There was no post-evaluation meeting.” She would take inventory of the class, without really making an effort to understand the topic or stage of the learning cycle, and follow up with “a lot of real critical comments in the write up.”

David believes that his negative evaluation, which “showed one-third needs improvement,” was personal. He admitted that he and his evaluator had a “mutual dislike” for each other, but had reached a truce several years prior. He described her as “extremely argumentative,” stating, “She really does not like me.” During the year of his negative evaluation, David and his supervisor once again clashed, with his evaluator “taking things to a personal level and us[ing] the evaluation tool to make a point. A political point.” The stark contrast between David’s ratings that year and those from the previous year, which were 100% proficient or advanced, appears to confirm his suspicions.

Amy also sees the potential for the tool to be used to intentionally favor undeserving teachers as well. Amy disclosed that she feels “in certain towns, towns that are very close, without at least one part being an outside evaluator, it will never be a true evaluation that you could actually trust.” Even with the use of the rubric, Amy sees evaluation as largely subjective and somewhat dependent on one’s relationship with their evaluator. For example,

if I’m the evaluator and I love this teacher… but she doesn’t upload evidence for the books element, then it’s very clear that I need to speak to her about that, or say
something negative about that, or whatever. So it adds a small step into that process, but then what usually happens, as I understand it, is they’ll just go and say, “Hey, you need to do this, you need to tweak this and you’ll be all set.” Evelyn concurred, “I get the impression that they’re using it as a way to justify what they already think about us.” Amy commented that she felt as if her evaluator “was kind of a little bit, ‘Wink, wink, you’re gonna be fine.’” David recalled how his experience in one district, yet two different schools, yielded a very different outcome despite his evaluators’ use of the same rubric, same standards, and same indicators. For David, it is evident that a “different evaluator” had the greatest impact. Matthew commented that his vastly dissimilar evaluation experiences in two separate districts signify that evaluation “is dependent on the evaluator, even though it shouldn’t be. It should be dependent on the tool being used itself.” Evelyn acknowledged that despite the detailed rubric mandated by the state, the evaluation itself is “subjective.” She admitted that she has to submit “very little evidence and they’ll give me exemplary,” while some of her colleagues input considerably more evidence in order to get the same rating. The presence of the educator evaluation rubric should improve objectivity; however, Amy thinks, “at the end of the day… it’s just more of the same.” Whether the evaluation tool is being used for teachers’ benefit or to their detriment, participant responses indicate that the personal relationship between the evaluator and the educator does impact the evaluation outcome.

Additionally, participants conveyed that it seems as though administrators are using evaluation to incentivize teachers into complying with specific initiatives. For example, Evelyn explained that
they’re trying to get us to do class websites, but they can’t really force us to do it, it’s not in our contract, so they’ve definitely picked the different teachers and given them needs improvement for parent communication even though maybe they’re doing a weekly email home or they’re doing something, and then the principal recommends, well if you want to get proficient, you should probably make a website like this other teacher has.

Theresa also expressed the belief that school leaders are using evaluation to strongly encourage certain behaviors in their teachers, which becomes overwhelming since then those behaviors come to be expected as typical. Of her principal, Theresa realized, You just can’t do all the things she wants you to do… for example, she wanted the parent involvement, she wanted the people to come in three times so she expects that now to be part of my teaching. So now I have to add more to it.

David believes that, in his case, the teacher rubric was used by his previous evaluator punitively because she viewed him as a “giant pain in the ass” for questioning school policy changes that were “dropped in our lap” or “contradict[ed] a contract.” He expressed concerns about using evaluation “politically to make statements” or “push teachers out in the way that they’re doing.”

**Conclusion**

The participants, who were purposively selected based on their common experience as middle school teachers of state-tested curriculum areas, represent large and midsize, urban and suburban districts. An analysis of their experiences related to educator evaluation exposed several major themes. Participants see the current evaluation model as top-down and accountability driven, with district and school leaders
enacting policies that compel teachers to be more liable for student achievement growth. Participants acknowledged that the educator evaluation tool, and especially the goal setting requirements and teacher rubric, provides structure to the system of evaluation, although the vast amount of documentation is viewed as overwhelming and time-consuming for both evaluators and teachers. Participant experiences reveal that evaluation procedures and expectations are inconsistent among and within school districts, especially regarding the educator performance rating, despite the presence of a common rubric and common language. The evaluator/teacher relationship is instrumental to the overall outcome of the evaluation, and in places where evaluation is perceived as more threatening and less clear, participants seek out relationships with coworkers for support and stability. Participants believe that the evaluation system is manipulated for political reasons to retain or dismiss teachers, sometimes arbitrarily, and as a coercive mechanism for encouraging educator compliance with district or school-wide policies or initiatives. In summary, participant views about the purpose of evaluation are not suitably aligned with many of the methods districts have employed to address state mandates related to the topic.
Chapter 5: Discussion of the Research Findings

This chapter will revisit the problem of practice, review the methodology used to conduct the study, and discuss the major research findings. The work will also consider the findings in relation to the theoretical framework and the key concepts presented in the literature review. Finally, the chapter will include a discussion of the implications for practice, limitations, and significance of the study in the broader context of the research problem.

Revisiting the Problem of Practice

In early 2012, the Massachusetts Department of Elementary and Secondary Education released a newly revised educator framework for evaluation, a requirement under the Race to the Top grant program. The Model System for Educator Evaluation aimed to more clearly identify an individual teacher’s contributions to student learning by rating teacher performance on four overarching standards and assigning teachers an Impact on Student Learning rating based on student test score data. School districts across Massachusetts were afforded a degree of control over how to adopt or adapt the educator evaluation regulations, yet all districts were expected to report Educator Performance Ratings and Impact on Student Learning Ratings to DESE (Massachusetts Department of Elementary and Secondary Education, 2017). The Model System signaled a major departure from most districts’ evaluation systems at the time, which were primarily based on observation and occurred somewhat infrequently. Additionally, the reporting of educator performance based on a rating and the reliance on student test scores to rate teacher performance based on low, moderate, or high growth was a novel concept. Whereas schools had been accountable for reporting on student performance
since the passage of No Child Left Behind, the reporting of teacher performance was
different, generating anxiety among educators as they sought to make sense of the new
regulations.

The adoption and implementation of new evaluation regulations that rate
educators based on performance standards and student achievement occurred with
minimal input from teachers. While unions may have had an opportunity to collectively
bargain over the specific language used in evaluation in some districts, most teachers
were compelled to acquiesce, even if they questioned the fairness or reliability of these
new evaluation methods. Teachers of state-tested subjects were potentially more
vulnerable, since standardized testing data could be used to determine their Impact on
Student Learning Ratings. Therefore, this study sought to understand how middle school
English, math, and science teachers experienced changes to evaluation and the effect of
these changes on professional practice.

Review of Methodology

The purpose of this study was to understand how middle school English, math,
and science teachers have experienced changes to evaluation. Middle school teachers
were selected based on content specialization, the heterogeneous grouping of students at
that level, and the standardized testing schedule, which mandates ELA and math testing
for students in grades six, seven, and eight and science testing for grade eight. The study
sought to answer the following question:

- What are the experiences of educators whose practices have been shaped
  by changes to the system of teacher evaluation?
To answer this question, the researcher relied on a qualitative Interpretative Phenomenological Analysis approach to capture the experiences of six participants who were selected through purposive sampling. Participants with a minimum of five years of teaching experience were chosen based on their subject taught, years of experience, and years in their current role. The researcher applied a social constructivist paradigm to effectively highlight the unique experiences of the participants, engaging with them interactively to co-construct meaning from their experiences. As a teacher and evaluator, the researcher understands that the participant experience depends on factors that are outside of the teacher’s control, such as the district leadership, building leadership, working conditions, and student demographics.

**Discussion of Key Findings**

The findings of this qualitative study were the result of careful analysis of the data, derived from semi-structured one-on-one interview transcripts and the researcher’s notes and personal memos and reflections. The researcher relied on member checking at multiple points during the analysis process, allowing participants to view interview transcripts and comment on the researcher’s analysis and interpretations. Additionally, the researcher met regularly with a small group of professional colleagues to discuss the work and assist with peer review of the research.

Data analysis using first and second cycle coding resulted in the following emergent themes:

**Accountability driven.** Participants recognized that the Model System of Educator Evaluation is a top-down initiative, required due to changing state regulations pertaining to the evaluation of teachers. As teachers of state-tested curriculum areas,
participants readily identified a link between test scores and evaluation. Some participants theorized that the rating of school districts based on test scores and the rating of educators based on standards and student learning impact resulted in increased scrutiny of individual teachers and a need to justify low test scores with non-proficient educator performance ratings. Those participants who teach high-needs student populations or in districts with a history of low test scores indicated that they felt increased pressure from administrators to raise achievement levels, even during the hold harmless phase. The result in these districts has been a climate preoccupied with test preparation. Participants recognize a trend over the past several years to shift more responsibility for student learning onto teachers and away from the students themselves.

**Structured.** In discussing the Model System of Educator Evaluation, participants highlighted some of its advantages, describing the overall evaluation process as more comprehensive and representative of their teaching than a single observation. Participants appreciate the clearly defined performance expectations included in the rubrics. Those who rely on evaluation management software report satisfaction with being able to upload documents that showcase what they do in the classroom in order to address each of the standards. Goal setting was identified as a significant component of the evaluation process, and those participants who established individual goals reported higher levels of satisfaction with the process than those who established team goals. The creation of team goals is complicated by the placing of teachers on different years of the evaluation cycle, the potential for different evaluators, and the use of team data for measuring student learning. Each of these factors makes it more difficult to compare
individual teachers, demonstrating why the rating of educator performance is problematic and potentially subjective.

Some participants discussed reviewing test score data in order to develop goals based on prior testing. Those who review test data reported doing so out of curiosity, and most admitted to not fully understanding how to interpret the data or how district leaders use this information. Additionally, teachers described feeling hesitant to establish goals that were too challenging, especially if being compared to teachers who created easier goals.

**Time consuming.** Participants were in agreement that the new evaluation system takes a considerable amount of time for teachers and evaluators due to the documentation required. Several participants explained needing to provide evidence for every single indicator, as well as the student learning and professional practice goals, while others simply felt the need to submit enough evidence to thoroughly capture their teaching. Even for teachers in districts that identified certain focus standards, evaluation procedures were described as time consuming and overwhelming. Participants also recognized the increased workload on their evaluators, who appear inundated with observation requirements and paperwork. A common feeling among participants was that it was almost impossible to get evaluation right and make the process completely worthwhile and authentic under the current system due to the increased demands placed on teachers and evaluators.

**Inconsistent.** State regulations mandate a five-step process for evaluation using common rubrics and common language, yet district leaders were given the opportunity to adopt, adapt, or revise their systems of evaluation to comply with the regulations and
establish guidelines for implementation. As a result, evaluation procedures vary among and within school districts. Participants expressed confusion over the criteria used to rate teachers as Exemplary, Proficient, Needs Improvement, and Unsatisfactory. With the Proficient rating encompassing a wide performance range of barely proficient to nearly exemplary, participants believe that the rating may not be truly indicative of teacher effectiveness. Additionally, one participant communicated disappointment in his district’s decision to rate all non-professional teachers as Needs Improvement, while another was discouraged by his superintendent’s decision to not let anyone be rated as Exemplary. Participants appeared uncertain about how the ratings are determined and who reviews that information, describing a lack of clarity from district leaders.

Participants shared a common belief that that evaluation procedures differ within their schools. This assertion was made by several participants due to their experiences with different evaluators and when comparing evaluation expectations with colleagues. Whereas some evaluators prefer long, thoughtful reflections, others request minimal evidence. While some evaluators write detailed narratives, others provide very little written feedback. The inconsistency in evaluation procedures is exacerbated when school administrators and district leaders fail to effectively communicate evaluation expectations to teachers.

Relationship dependent. Participants discussed the relationship between the evaluator and teacher as being a significant factor of evaluation. The evaluator is responsible for setting the tone for the evaluation process. Participants who had a positive personal relationship with their evaluator were more satisfied with their evaluator’s performance. Participants described effective evaluators as those who were
fair, honest, took the job seriously, and ideally had content expertise in the subject for which they were evaluating. Several participants stated they felt it was important to get to know the evaluator to anticipate what he or she wanted in order to make the relationship mutually beneficial.

In schools where evaluation has been viewed more negatively, participants credited their resilience to the strong relationship networks forged with colleagues. Peer collaboration and support assists teachers with managing evaluation requirements and difficult evaluators. One participant discussed coaching colleagues who were facing negative evaluations, while another discussed the importance of collegiality and collaboration for creating stability in an otherwise unstable work environment that is susceptible to ever-changing policies and expectations. The degree to which participants depend on colleagues for support is inversely related to how positively they view the evaluation process.

**Political.** Participants were unanimous in their assertion that the evaluation tool is being used by evaluators to confirm what they already think about teachers. In some cases, participants believe the educator performance rating is being arbitrarily assigned to teachers based on superficial reasons, like personality or whether or not they get along well with the administration. Participants were in agreement that the performance rating is not necessarily an accurate indicator of teacher effectiveness due to the tendency for evaluators to manipulate the rubric to fit their predetermined ideas about the teacher. Additionally, some administrators appear to use evaluation to coerce teachers into implementing certain initiatives that otherwise could not be required in exchange for a
favorable rating. Participants’ experiences seem to confirm that the new evaluation regulations, while more structured and comprehensive, do not guarantee objectivity.

**Discussion of the Findings in Relation to the Theoretical Framework**

Hall and Loucks’ (1978) Concerns-Based Adoption Model is a suitable framework for making sense of teachers’ experiences since it focuses on how people respond to change. Public school teachers have had to comply with educational regulations, adopted in 2012, that require them to participate in the five-step evaluation cycle, consisting of goal setting, performance standards, and ratings; however, teacher input involving these regulations has been minimal. Therefore, the CBAM is a valuable tool for identifying how participants have responded to changes in evaluation.

CBAM assumes that change is a process, the process is personal, and success depends on the ability of individuals to adapt their practices and embrace the change in order to bring about systematic change (Horsley & Loucks-Horsley, 1998). The framework places teachers on a Stages of Concern continuum based on their feelings and motivations related to the change (Hall & Loucks, 1978). Once their degree of concern regarding the change is determined, teachers’ Levels of Use can be identified (Gundy & Berger, 2016). The levels of use depend on multiple factors, including “implementation assistance, time and experience with implementation, and administrative pressure and support” (Anderson, 1997).

All of the participants in this study had been working under the current evaluation system for several years and during that time were made aware of the change and given varying levels of information regarding the change. The placement of participants on Stages 2, 3, or 4 is dependent on the degree of communication they have received from
district and building leaders, as well as the support they have received in implementing
the changes. The next section will address the Stages of Concern and Levels of Use for
each participant.

**Evelyn.** Evelyn’s experience in multiple school districts and as a practicum
supervisor has given her confidence to implement the evaluation changes. She
understands that she needs to take inventory of her work environment in order to
effectively manage evaluation components. Evelyn is aware that her teaching assignment
drives goal development and she recognizes the role her relationship with her evaluator
plays in her evaluation outcome. Evelyn is not only managing evaluation, she is invested
in improving the process to benefit student learning. While Evelyn would appreciate
being able to collaborate more often with colleagues, she also realizes her position as a
middle school science teacher confines her somewhat. Evelyn’s ability to progress
beyond Stage 4 is limited by circumstances out of her control. Evelyn appears relatively
comfortable with evaluation procedures, as they have become routine, yet her insistence
on developing student-learning goals for her classroom that align with district goals
demonstrates a commitment to refining the process.

**Theresa.** An experienced educator, Theresa’s knowledge about evaluation is
extensive. Like Evelyn, Theresa’s confidence in her teaching ability has helped her to
successfully progress along the Stages of Concern continuum when meeting new
evaluation requirements. Theresa’s strong desire to improve student-learning outcomes
for her students and her interest in collaborating with colleagues places her between
Stage 4, Consequence, and Stage 5, Collaboration, on the Stages of Concern scale.
Theresa’s level of use appears somewhat dependent on her principal’s insistence that
teachers incorporate specific actions into their routines in order to achieve proficiency in those areas. While Theresa has acquiesced and met those demands, she also described feeling overwhelmed at the number of new initiatives that teachers are expected to absorb into their regular practice. Theresa’s level of use is best characterized as fluid, moving between routine and integration depending on how much value she places on the specific requirement. For example, bringing parents into school, which was an expectation of the principal, was something that Theresa complied with but did not necessarily see as essential, and thus could be seen as routine. Using data specifically for the purpose of increasing student achievement, however, demonstrates adeptness at integrating the change.

Matthew. As the participant with the least teaching experience, Matthew has a solid handle on evaluation procedures due his experience working in several different school districts and his C.A.G.S. training, which he credits with helping him understand evaluation terminology and best practices. As the Team Lead teacher, Matthew assumed additional responsibilities and subjected himself to increased scrutiny as the one responsible for overseeing curriculum and assessment for his team. Matthew appeared at ease with evaluation requirements and viewed personal reflection and collaboration with his evaluator as strengths of the evaluation system. His frustration is mainly with his superintendent’s policy concerning educator performance ratings as opposed to the Model System of Evaluation itself. Matthew has a sound understanding of how to use student data to inform his practice, yet recognizes that his district’s implementation of the Model System of Evaluation stymies his potential for reaching Stage 6, Refocusing, and in turn, his levels of use is limited by district policy. The district’s practice of viewing
student learning data holistically rather than attributing gains to individual teachers
discourages Matthew from progressing beyond routine and refinement since there is no
perceived incentive to do so.

Amy. Confident in her ability to use TeachPoint to upload evidence and
demonstrate her impact on student learning in measurable ways, Amy registers at a Stage
4, Consequence, on the Stages of Concern framework. She describes the evaluation
process as easy and comprehensive and is pleased that the new system of evaluation
encompasses more elements of teaching than simply classroom learning. Amy is
constricted by her work environment, however, collaborating minimally with her
colleagues and receiving limited direction from supervisors; therefore, she does not
receive the current level of support required to be able to effectively collaborate or
refocus. Amy continues to refine her habits, using student test scores to develop and
assess her goals. Her ability to reach the integration and renewal levels of use depends
on additional support from district and building leaders.

David. David’s skepticism regarding the Model System of Evaluation is largely
due to his negative experience with his previous evaluator. He also expressed concern
with standardized testing in science, which is only done in fifth and eighth grade and
assesses students on concepts that may not be developmentally appropriate. David’s
experience with his current evaluator has been positive, however, and this has helped him
manage evaluation components, which places him at Stage 3. As far as levels of use are
concerned, David appears to be at the mechanical level, where he is figuring out the
logistics of how to review student achievement data using tools like MasteryConnect in
order to impact teaching and curriculum development. Movement beyond his present
stage and level is contingent on a continued relationship of trust between David and his
evaluator and the district’s offering of training in the review of student achievement data
using programs available to teachers.

Jack. As a teacher in a district labeled as underperforming, Jack is the least
comfortable with the Model System of Evaluation, specifically due to concerns with
assigning teachers a rating based on their student performance on standardized tests. Jack
passionately advocates for his students, many of who are considered underprivileged.
Expressing strong reservations over the use of student test scores to evaluate teacher
performance and his district’s current practice of labeling non-professional teachers as
Needs Improvement, Jack registers at a Stage 2, Personal, on the Stages of Concern
framework. Jack’s level of use falls between mechanical and routine. He recognizes
what he needs to do to comply with the requirements and his actions have become
routine, yet Jack still looks to his principal for guidance on goal setting and is not
convinced of the benefit of the evaluation changes to his students.

Summary. Participants who are employed in districts with Level 1 or Level 2
ratings have progressed further along the Stages of Concern continuum than those who
teach in districts with lower ratings, as determined by student standardized test scores.
Furthermore, those participants who have received clear instructions on evaluation
expectations and who feel supported by their evaluator have demonstrated higher levels
of competency when it comes to goal setting and gathering evidence. All of the
participants are restricted by the top-down nature of the evaluation system since teachers
remain uncertain about how educator performance ratings are determined and how this
information is used and reported. Therefore, teachers are hesitant to fully embrace the
Model System of Educator Evaluation because they do not have complete confidence in those who are executing it to use it as it was intended.

**Discussion of the Findings in Relation to the Literature Review**

The Massachusetts system of teacher evaluation was revised in 2011 to satisfy the educational requirements of the American Recovery and Reinvestment Act. As a result, the Model System for Educator Evaluation was developed with a goal of closing the achievement gap between high- and low-performing students by more effectively determining how individual teachers contribute to student learning. Additionally, Massachusetts adopted the Common Core State Standards as a framework for teaching core content areas. In turn, school districts modified curriculum as necessary to align with the revised frameworks, and district leaders began the process of implementing the new evaluation system.

Among the concerns highlighted by educational reformers in their quest to increase educator accountability is that traditional observation only evaluations did little to improve learning and too many teachers were rated proficient (Ambady & Rosenthal, 1993; Attinello et al., 2006; Jacob & Lefgren, 2008; Medley & Coker, 1987; Strong et al., 2011). The findings from this study suggest that the educator performance ratings have not made the process for distinguishing the effectiveness of those teachers in the middle any easier. Participants intimated that they were frustrated by reluctance on the part of evaluators to differentiate between high proficiency and low proficiency and the inability to communicate this information to DESE using the model rubrics. Furthermore, the new evaluation system is considerably more work for teachers than previous methods of
evaluation, and there is little perceived value if teachers are unable to affect their performance ratings.

Efforts to more closely correlate teacher effectiveness with student performance have led to the proliferation of value-added models for educator evaluation (Ballou et al., 2004; Braun, 2005; Hill et al., 2011; McCaffrey et al., 2003). In Massachusetts, the proposal to evaluate teachers partly on student test scores is problematic since participants indicated that they do not know how to accurately interpret data from state testing reports or use data in their own classrooms to revise their teaching and curriculum. This reinforces Collins’ (2012) findings that reveal teachers and administrators did not know how to interpret data from value-added models. Participant experiences also suggest that evaluators in lower-performing districts may be rating teachers based on past student test scores to justify the district’s lower scores rather than evaluating teachers based on their own students’ actual learning. This tail wagging the dog approach for linking teacher evaluation to student test scores undermines the credibility of the evaluation model if it is truly being used to fit an existing narrative rather than inform future practice.

The literature examined the ways in which curriculum has been impacted by standardization and accountability reforms. Researchers determined that standardized testing requirements often drive curricular offerings (Polesel et al., 2012). Participants confirmed that courses are developed based on testing requirements, with the goal of raising test scores at the forefront of district and school-based decisions. This is especially the case in districts with a history of lower standardized test scores. Furthermore, student assignments are not random, and this complicates the evaluation of
teachers based on student learning. Several participants reported being assigned certain high-risk student populations due to their reputation for success with these groups of learners. This is consistent with the research that found students are not randomly assigned to teachers (Rothstein, 2009). Furthermore, those participants who began teaching prior to the enactment of No Child Left Behind all commented on the shift away from student ownership of learning and toward greater educator accountability. This coincides with existing research that found that high-stakes testing negatively impacted student learning and motivation (Amrein & Berliner, 2003).

**Implications for Practice**

Reliance on a qualitative, IPA research approach was beneficial for conveying how teachers have been trained in, and have experienced, the new evaluation tool that rates their performance on student learning and professional practice goals progress, as well as on standards and indicators of effective teaching. The research provides information about what areas of the evaluation tool need clarification and sheds light on the potentially harmful aspects of relying on performance ratings for evaluation.

The Concerns-Based Adoption Model framework is useful for understanding the ways teachers have adapted their practices to align with components of the evaluation rubric. The findings demonstrate that teachers’ full proficiency in the use of the evaluation tool is stymied by districts’ inability or unwillingness to make the process of evaluation authentic so that teachers are willing to invest the time and energy necessary for professional growth. Policymakers and members of the Massachusetts Board of Elementary and Secondary Education could encourage buy in from educators if they were to clarify evaluation expectations for district leaders. In turn, district leaders could
provide initial training for new evaluators and regular calibration activities for all evaluators to encourage healthy communication between evaluators and educators, as well as ensure consistency in evaluation procedures. If this occurs, teachers should be able to demonstrate effectiveness with a minimal amount of supporting evidence so that the educator rating reflects the quality of the teacher’s efforts rather than the amount of time spent documenting those efforts.

Evaluation has the potential to empower educators when implemented in a way that gives teachers a feeling of control over the process. The findings reveal that many teachers want to review and analyze data from standardized tests to improve their instruction in the classroom, but are not certain how to do so. Professional development training in the use of student learning data to inform practice would equip teachers with the knowledge to assess and revise curriculum and instruction to effectively address gaps in student learning.

As a department head responsible for the evaluation of teachers, the researcher will mindfully implement practices to encourage personal growth and reflection among the teachers she supervises in an effort to make evaluation valuable and authentic. The importance of the evaluator relationship cannot be understated, and therefore the researcher will make an attempt to foster relationships based on trust and respect. As an evaluator, the researcher will engage in regular conversations with teachers and encourage the development of authentic student learning and professional practice goals. The researcher will visit classrooms regularly to maintain a pulse on teaching and learning in the schools and engage in conversations with teachers about how to support them in their continued development.
The researcher will also continue to discuss evaluation procedures with other department heads, building leaders, and district leaders in order to strive for more consistency in evaluation practices. The findings from this study suggest that teachers are well aware of inconsistencies and do discuss their evaluation experiences and performance ratings with colleagues. Policies that allow for transparency should be pursued and evaluation should not be used as a coercive mechanism or politically, to artificially bolster or diminish a teacher’s accomplishments or contributions.

**Limitations**

The study is limited by several factors. First, the sample size is small, consisting of only six participants. The participants represent only the state-tested subject areas of English, math, and science, as these are the subjects most significantly impacted by curriculum and evaluation changes. Therefore, the findings do not consider teacher experiences in alternative subjects or grade levels, which limits generalizability. Despite these limitations, efforts were made to choose participants from school districts of different sizes that serve a range of student populations in order to obtain a range of views and opinions. Participants were eager to share their personal experiences related to evaluation.

**Significance of the Study**

In the wake of recent reforms that place greater responsibility for student learning on individual teachers and use student growth data to determine educator performance, public school teachers in Massachusetts have had to comply with changes to regulations pertaining to evaluation. The findings from this study reveal that changes to evaluation have impacted teachers both positively and negatively. On the one hand, teachers
appreciate the thoroughness and transparency of the new system and are pleased with the ability to submit evidence that comprehensively addresses their work as educators, rather than an observation-only based evaluation. At the same time, they feel burdened and overwhelmed by the time required to effectively document they are doing their jobs. Efforts to streamline procedures by relying on evaluation management software or encouraging team goals have not necessarily made the process less of a burden.

Participants view the evaluator relationship as a crucial component of the process. The findings indicate that evaluator training would be a worthwhile investment for districts in order to ensure that evaluators are adequately supporting the professional growth of their teachers. Research into the experiences of evaluators would be a useful endeavor for understanding the evaluator perspective. Additionally, many teachers are not aware of how to interpret student test score data, and training in this area would enable them to better revise curriculum and assessments to meet student learning needs. The findings also reveal that there are inconsistencies regarding evaluation expectations among and within districts. The Department of Elementary and Secondary Education, in its upcoming revisions to evaluation guidelines, should more clearly define the criteria for each of the educator performance categories and provide exemplars for teachers and evaluators to ensure consistency in procedures. Ultimately, the findings are important in the consideration of how revised educator evaluation guidelines impact the gatekeepers of educational reform.

Conclusion

Educational reform initiatives in recent decades have led to the enactment of policies that have expanded the federal government’s influence in the realm of public
education. Legislation that called for greater school accountability was eventually superseded by efforts to revise educator evaluation mechanisms to better reflect student achievement (Amrein-Beardsley & Collins, 2012; Mangianti, 2011). As a Race to the Top recipient, Massachusetts amended its education regulations, and superintendents were informed that their districts needed to modify their evaluation methods to comply with the regulations. As schools started to implement new evaluation regulations, teachers across the state began developing goals and collecting evidence.

The research study gives a voice to the teachers who have been significantly impacted by changes to evaluation procedures. Participants spoke of a desire to have a transparent and well-designed system of evaluation that improves and develops teachers as professionals. Their experiences under the current system, however, do not align with this vision. Rather, teachers are spending a considerable amount of time to prove they are doing their jobs, with little perceived value for themselves or their students. The recent DESE decision to eliminate the Impact on Student Learning rating and instead roll student achievement into Standard II signifies that the state’s education leaders, in consultation with union leaders, recognize the potential issues with basing teaching effectiveness on student test scores. The continued rating and reporting of teachers in their evaluations, however, persists, and the findings from this study reveal that these ratings are subjective in many cases, and thus not an accurate measure of teacher effectiveness. Future changes to evaluation procedures should take into account the opinions of the teachers themselves, since they will be the ones to experience the most significant effects, prior to implementation.
References


doi:10.3102/0013189X08316420


Collins, C. (2012). *Houston, we have a problem: Studying the SAS education value-added


Diamond, J., & Spillane, J. (2004). High-stakes accountability in urban elementary
schools: Challenging or reproducing inequality? *The Teachers College Record*, 106(6), 1145-1176.


teacher evaluation policies promoted by race to the top. NCEE Evaluation Brief. NCEE 2014-4016. National Center for Education Evaluation and Regional Assistance.


Horn, C. (2003). High-stakes testing and students: Stopping or perpetuating a cycle of failure? Theory into Practice, 42(1), 30-41.


doi:10.3102/0162373711422377


Massachusetts Department of Elementary and Secondary Education. (2016). Quick reference guide: Opportunities to streamline the evaluation process.


Reback, R., Rockoff, J., & Schwartz, H. L. (2013). Under pressure: Job security, resource allocation, and productivity in schools under NCLB.

Reiners, G. M. (2012). Understanding the differences between Husserl’s (descriptive) and Heidegger’s (interpretive) phenomenological research. *J Nurs Care, 1*(5), 1-3.


Scherrer, J. (2012). What's the value of VAM (value-added modeling)? Value-added modeling is likely here to stay. Although VAM has some advantages over traditional teacher assessment methods, there also are shortcomings that should be noted. *Phi Delta Kappan, 93*(8), 58-60.


survey about student-teacher assignment. *American Education Finance and Policy 37th Annual Conference, Boston, MA,*


improve the distribution of teacher quality? A simulation of alternative policies.


