EXAMINING TEACHER PERCEPTIONS OF THE RELATIONSHIP BETWEEN EVALUATION POLICY AND TEACHER PRACTICE IN A NORTH CAROLINA SCHOOL SYSTEM

By

Amanda Marie Slaten Frasier

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Educational Policy- Doctor of Philosophy

2017

ABSTRACT

EXAMINING TEACHER PERCEPTIONS OF THE RELATIONSHIP BETWEEN EVALUATION POLICY AND TEACHER PRACTICE IN A NORTH CAROLINA SCHOOL SYSTEM

By

Amanda Marie Slaten Frasier

Examining the justification for current evaluation policy reveals that such policy rests on two assumptions related to the impact on the work of teachers: (1) evaluations are necessary because teachers need to be rated, sanctioned, or rewarded in order to be motivated to improve their practice; and (2) evaluations yield information that is useful for teachers to improve practice. Both assumptions have driven policy changes over time and carry implications for teacher classroom practice.

This mixed methods study examines how a state-wide standardized evaluation policy utilized in North Carolina affects the work of high school teachers in a single school district under varying school and individual conditions. Specifically, this study focuses on teachers who offer perspectives from varying combinations of the following school-level variables: status at a high or low evaluation condition school and status at a high or low evaluation effectiveness school, and the following individual variables: status as a Mathematics or English teacher, years of experience, and licensure level.

This dissertation tests the previously-stated assumptions about teacher evaluation and teacher work in a North Carolina school system in a to answer the following research questions:

(1) What, if any, role do reported school evaluation conditions and school evaluation status play in shaping teacher motivation, experiences with feedback, and work decisions related to teacher evaluation?

(2) What individual-teacher level factors are associated with differences in teacher motivation, experiences with feedback, and work decisions related to teacher evaluation?

Analysis of the whole sample demonstrated that teachers did not find evaluation to motivate performance or to provide useful feedback. Though quantitative differences between school locations were not found, there were qualitative differences in how evaluation was related to practice across sites. Differences were also found in the evaluation-practice relationship between teachers of different licensure levels and different levels of experience where those in the lower designation perceived a greater impact of evaluation policy. Finally, differences between the subject areas of Math and English were identified, but may have been influenced by the capacity of observers and specifically, a lack of subject area alignment between the observer and the classroom in English, such alignment was present for some of the Math teachers in the study. Therefore, it is important to examine the context of evaluation, particularly the capacity of the administration that conducts evaluation.

The results of this study suggest that the characteristics and capacity of an observer do matter in how the observation protocol is interpreted and implemented. Additionally, the evaluation climate and culture, or evaluation scenario of a school, may also influence the ways in which teachers find evaluation motivating and how teachers approach feedback from evaluation. The results of this study provide insight into the relationship between teacher evaluation and classroom practice, an area that has previously been under researched despite the impact other high-stakes accountability policies have had on teaching practices and the teaching workforce.

For my mom

“A mother is she who can take the place of all others but whose place no one else can take.”

ACKNOWLEDGEMENTS

In all the time it took me to research and write this dissertation, I never expected this part to be the hardest to write. I would first and foremost like to thank Dr. Michael Sedlak for recruiting me to the Educational Policy program at Michigan State University, for serving on my initial guidance committee, and for putting together a 2012 cohort of colleagues which offered immense professional and personal support to me throughout the program. The program provided me with invaluable opportunities to travel, research, and network which graduate students in other programs can only dream of. So, I am forever indebted to this program for the professional and personal development that I have received.

I would also like to thank some fellow graduate students who provided feedback on earlier iterations of this work. Specifically, I must mention Alyssa Morley, Iwan Syrahil, Sarah Galey, and Jihyun Kim.

Additionally, I must thank my fantastic committee. Dr. Anne-Lise Halvorsen became my academic advisor in my second year of the program and as such has been with me through a myriad of both personal and professional ups and down. We have worked together on several projects and I have learned an immense amount about academic life from her. Dr. Peter Youngs approached me about working together when he was transitioning to a position at another university. I am so grateful that he stuck by me and offered his expertise despite the fact that he was several states away. It says a lot about Peter that he was willing to keep working with students at another university despite not having any formal obligation to do so. Dr. Corey Drake and I met together while I was working as a graduate assistant on a project, and she agreed to become a member of my dissertation committee just a year prior to my dissertation when I

realized that I would be examining some issues in which she had professional interests. All three of the above mentioned committee members helped me through a period of my life where I very much was struggling. The second year of my graduate school program I was giving a final exam presentation in Peter's class when my phone rang and I found out that my mother was on life support. She died less than 48 hours later, leaving me to deal with not only the emotional ramifications of her death, but the legal, physical, and practical as well. All three of these people continued to work with me, guide me, and mentor me at a time when I would otherwise have very well quit everything. Sometimes I only got by because of the immense sense of obligation I felt to these people. The final member of my committee, Dr. Madeline Mavrogordato joined my dissertation committee later on and really helped shape the methods of this study as well as my understanding of school leaders' roles in implementing evaluation policy. I have especially enjoyed sharing a personal connection with Maddy over our mutual love of horses, dressage, and Chick-Fil-A. These four individuals have helped shaped this work and shape me as a professional. Thank you for your time, your expertise, and your belief in me.

I also must extend my appreciation to the anonymous teachers who participated in this study. I wish there was more I could have given them for their help and their dedication to the profession.

Finally, I have to thank my family. I came from a family where college was never discussed as an option, and despite being a high achiever in high school, I almost did not attend. I have to thank my good friend, Jeff, for helping me enroll in community college. He is as much family to me as anyone and without him I may still be waitressing at a Denny's. I have to thank my sister and grandmother, neither of whom always understood what I was doing or what I have been up to, but who have supported me with their whole hearts throughout my entire education.

(I promise, Aimee, this is the last graduation of mine you will ever be dragged to.) My husband, Chad, earned his PhD a few years ago and until then I had no idea that one could earn an advanced degree without having wealthy parents to pay for it. His experience prompted me to seek out my own doctoral program and he has done his best to support me in the same ways I supported his education.

And finally, I need to thank my mother. How I wish so badly that you were here to see the conclusion of this journey! The truth is that the loss of you was as much of my graduate program as anything else. Over four years ago I was headed to my first professional conference in Europe and you dropped a bomb on me by telling me over the phone that you were having surgery to remove a cancerous kidney. You told me to get on the plane and I did. And I did it again and again. Losing you was the reason I wanted to quit and the reason I could not all in one. As I told you on the day you died, everything I am and everything I ever will be is because of you and I reflect on what you have taught me and on our relationship every single day. As I write this I am eight months pregnant and my only hope is that I can have the same influence on your grandchild that you had on me. Thank you for being my mother, for being my strength, and for always believing in me. This is for you. This is because of you. I love you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

KEY TO ABBREVIATIONS

AIG	Academically Intellectually Gifted
ARRA	American Recovery and Reinvestment Act
AYP	Annual Yearly Progress
CCSS	Common Core State Standards
EC	Exceptional Children
ELA	English Language Arts
ELL	English Language Learner
EOC	End of Course
ESEA	Elementary and Secondary Education Act
FARPL	Free and Reduced Price Lunch
MET	Measures of Effective Teaching
NC	North Carolina
NCFE	North Carolina Final Exam
NCES	National Center for Educational Statistics
NCLB	No Child Left Behind
PDP	Professional Development Plan
PLC	Professional Learning Community
RttT	Race to the Top
SWD	Students with Disabilities
TWC	Teacher Working Conditions
VAM	Value Added Model

CHAPTER 1: Introduction

Current teacher evaluation policies have emerged from policymaker critiques that previous systems of evaluation did not accurately identify the effectiveness of teachers and that many teachers were often rated as high performing. One example of such a critique is found in the Measures of Effective Teaching (MET) project, sponsored by the Bill and Melinda Gates Foundation, which began in 2009 and is the largest study of teacher evaluation to date. In justifying the project's worth, the Gates Foundation described previous evaluation schemes as "not providing the information needed to close the achievement gap. Despite 40 years of research pointing to huge differences in student achievement gains across teachers, most school districts and state governments cannot pinpoint what makes a teacher effective or identify their most and least effective teachers" (Bill & Melinda Gates Foundation, 2010, p. 2). The justification provided by the Gates Foundation identifies schools and districts as ineffective evaluators of teachers.

Furthermore, studies have found that large numbers of teachers have rated highly across states. For instance, the New Teacher Project study found that for evaluation systems with only two ratings, "satisfactory" and "unsatisfactory," 99% of teachers earned a satisfactory. In evaluation systems with more than two ratings, 94% of teachers received one of the top two ratings and less than 1% were rated unsatisfactory (Weisberg, Sexton, Mulhern, & Keeling. 2009). The Weisberg et al. study termed this top-heavy sort of assessment as "the Widget Effect." Aside from ranking teachers inaccurately, another criticism of local based evaluation systems is that the perfunctory nature of evaluation does not provide meaningful feedback to improve practice. These critiques have led to reform in evaluation, often centering policy at the state level rather than the district level, and typically including both standardized observation measures and growth data based on student performance.

Examining the justification for current evaluation policy reveals that such policy rests on two assumptions related to the impact on the work of teachers: (1) evaluations are necessary because teachers need to be rated, sanctioned, or rewarded in order to be motivated to improve their practice; and (2) evaluations yield information that is useful for teachers to improve practice. Both assumptions have driven policy changes over time and carry implications for teacher classroom practice.

Following the adoption of standardized observation protocols and value-added models (VAMs) meant to measure student growth by individual teachers, a large body of literature has examined both the technical aspects of evaluation (e.g., Baker et al., 2010; Bill and Melinda Gates Foundation, 2013; Corcoran, 2010; Glazerman et al., 2011; Goldhaber, Goldschmidt, & Tseng, 2013; Harris, 2009; Hill, Kapula, & Umland, 2011; McCaffrey, Lockwood, Koretz, & Hamilton., 2003; Raudenbusch & Jean, 2012; Rothstein & Mathis, 2013; Sanders & Horn, 1994) as well as the resource and infrastructure demands such systems place on schools and districts (e.g., Anagnostopoulos, Rutledge, & Jacobsen, 2013a; Mintrop & Sunderman, 2013; Thorn & Harris, 2013). The effective labelling and sorting of teachers into ranked categories is thought to be important because other research has been unable to identify the characteristics of effective teachers (Ballou & Podgursky, 1998; Boyd, Grossman. Lankford, Loeb, & Wyckoff, 2009; Darling-Hammond, Holtzman, Gatlin, & Heilig, 2005; Goldhaber & Brewer, 1997; Harris, 2009) or what type of preparation best prepares one for the classroom (e.g., Goldhaber & Hansen, 2010). What remains unclear from current research is (1) the effect that evaluation systems have on teacher classroom practice as individuals go through high-stakes individual evaluation cycles and (2) the extent to which teachers use feedback from evaluation to further guide classroom practice. This dissertation addresses this gap in the literature by examining teacher perceptions of the interaction between evaluation policy and classroom practice considering differences at the school- and individual- levels.

This study examines how a state-wide standardized evaluation policy utilized in North Carolina affects the work of high school teachers under varying school and individual conditions in the same school district. Specifically, this study focuses on teachers who offer perspectives from varying combinations of the following school-level variables: status at a high or low evaluation condition school and status at a high or low evaluation effectiveness school, and the following individual variables: status as a Mathematics or English teacher, years of experience, and licensure level (the latter two are linked to the number of evaluations a teacher receives).

For the purposes of this dissertation, I use responses from evaluation-themed questions from the 2016 administration of the biannual North Carolina Teacher Working Conditions (TWC) survey to identify school status as having high or low evaluation conditions based on site deviation from district averages. I define school status as a high or low evaluation school in a similar manner by using data from 2015-2016 from the Educator Effectiveness Database Section of the North Carolina School Report Card system. In both cases, data from 2015-2016 is used because that was the most current data available at the time of the study and represents the school year immediately preceding the study year. Additionally, I track varying characteristics of teachers such as their subject area certifications, years of experience (career status), and licensure level through survey responses. The rationale for and additional explanations of these definitions and methods will follow in a subsequent section.

Using the aforementioned variables, this dissertation tests the previously-stated assumptions about teacher evaluation and teacher work in a North Carolina school system to answer the following research questions:

(1) What, if any, role do reported school evaluation conditions and school evaluation status play in shaping teacher motivation, experiences with feedback, and work decisions related to teacher evaluation?

(2) What individual-teacher level factors are associated with differences in teacher motivation, experiences with feedback, and work decisions related to teacher evaluation?

These questions were answered in a mixed method study using a combination of quantitative data analysis of survey results and qualitative analysis of interview transcripts. Chapter 2 presents background information on teacher evaluation policy. The chapter starts with a brief history of teacher evaluation in the United States, particularly in relation to the larger movement to increase school accountability. Next is a legal review of how teacher evaluation policies are linked to an overall shift in governance over schools. Finally, I describe the historical, legal, and policy context of evaluation in the study state of North Carolina.

Chapter 3 provides a literature review and develops a framework to investigate questions about the relationship between teacher evaluation and teacher practice. The chapter starts with examining the two policy assumptions motivating teacher evaluation policy, namely teacher motivation and feedback use. There is a gap in current literature on the evaluation-practice relationship. Therefore, two related areas of literature are explored to anticipate how teachers may respond to evaluation policies in practice: teacher responses to external accountability pressure and to curriculum reforms. Because school administrators play a large role in how evaluation is implemented at the school-level, I also include a section on leadership capacity and evaluation. Finally, I present a framework for pursuing research on the relationship between teacher evaluation and teacher practice.

Chapter 4 includes the research design and methodology for this dissertation. The chapter includes a description of my sampling strategy, including methodology for calculating school-level Evaluation Condition and Effectiveness Scores and the selection of sites based on those calculations. Additionally, I delineate my three phases of data collection including: survey administration, preliminary focal interviews, and follow-up focal interviews. I then discuss how I analyzed data and established validity of my findings. Finally, I provide descriptions of the school system and the four school sites.

In Chapter 5, I use data from the entire sample of teachers to identify overall trends in how teachers perceive the practice and evaluation relationship. First, I present the results of questions which replicate the North Carolina Teacher Working Conditions survey, in which the original survey was used to calculate school-level Evaluation Condition Scores. Next, I compare teacher perceptions of the prior year to the study year. I then use the literature derived framework from Chapter 3 to examine both of the two policy assumptions of evaluation (as motivation and as a feedback tool). Additionally, I use literature from teacher responses to external accountability measures to identify similar responses fueled by the evaluation policy in this study. Finally, I identify teacher reform typologies using categories derived from literature on teacher responses to classroom reform.

Chapter 6 examines school-level differences across research sites. I initially hypothesized that schools with varying levels of effectiveness scores and varying evaluation conditions would perceive evaluation differently. However, the quantitative data showed no significant differences between schools. I then draw on interview data to explain the quantitative data and offer alternative theories. I also illustrate that despite the lack of statistical findings, there were stark differences in how evaluation affected teachers across schools as evidenced in the interviews. To do this, I present vignettes of each of the four schools and describe three evaluation scenarios which emerged from the interview data.

Chapter 7 investigates the individual-level teacher characteristics of: licensure, years of experience, and subject area to identify differences in how teachers of various characteristics perceive the relationship between evaluation and practice. Each of the three characteristics include a separate presentation of the survey data, interview data, and discussions of the characteristic. For the characteristic of subject area background, specific concerns around observation and testing are presented.

Finally, Chapter 8 will offer concluding thoughts on the dissertation. This will include implications for research as well as for the practice of evaluating teachers and for evaluation policy implementation at the school-level. Specifically, the following areas will be explored: leadership capacity, perceptions of evaluation validity, and altered teacher behaviors. Additionally, some of the possible unintended consequences of the evaluation policy in this study are discussed. Policy recommendations will be provided for how evaluation for both high and low stakes purposes may be reconciled to allow for more effective use. Limitations of this study are also discussed.

Overall, the results of this dissertation demonstrate that teachers do not find evaluation policy to be motivating or to provide feedback that is useful to changing practice. However, some unintended consequences of teacher evaluation policy emerge in patterns similar to what has been found in research on other external accountability measures. School-level results indicate that approaches to observation and testing are not standardized across sites, despite efforts to create a policy which is uniform. Additionally, the way in which school administration approaches the components of evaluation influences teacher perceptions, possibly impacting the success of the policy or further leading to unintended policy consequences that may negatively impact the teaching workforce and/or the work of teachers.

Additionally, some are differences demonstrated between groups of individual-level characteristics. For instance, differences between licensure and experience levels may be linked to the frequency of evaluation and the increased high-stakes for those who have lower levels of licensure or experience. The statistical differences in subject area may be related more to the conditions under which individuals are evaluated, particularly in regard to the capability of the evaluating administrator, rather than characteristics that are inherently linked to teacher subject area background. These findings suggest that despite attempts to standardize evaluation protocols, differences in school and individual contexts may result in differing evaluation experiences and differing relationships between evaluation policy and individual teacher practices.

At the time of this writing, there is a gap in the literature on how formal teacher evaluation policy is related to classroom practice. This is an important question to consider because evaluation, by definition, defines what is valued in whatever is being appraised. Additionally, such policies are touted by policymakers as being necessary to motivate teachers to do a better job and to provide feedback for them to do so. Therefore, it is important to consider whether formal policies do motivate and provide feedback to teachers, and if such policies do these things, then to consider in what ways teacher practice changes as a result? This dissertation begins to answer important questions around evaluation and practice as related to the study context. Such information is useful when weighing the costs and benefits of high-stakes teacher evaluation policies.

CHAPTER 2: Evaluation Policy Background

Evaluation is a process in which the characteristics of what is valued are identified and appraised. Traditionally, the evaluation process for teachers in the U.S. has been a local affair consisting of classroom observation and local personnel preferences, such as the teacher's ability to coach or teach certain subjects, with limited standardization among the protocols, frequency, or observers utilized (e.g., Tyack & Cuban, 1995). Cohen (2011) explained that in the past, "conceptions of teaching quality were tied to a teacher's years of education, degree attainment, and years of experience, none of which are closely related to the quality of work in the classroom" (p. 63). So, definitions of good teaching have often been determined at the local level. This created variety among evaluators (depending on preferences and experiences) and from site to site (depending on implementation and fidelity) due to variability in values among both individual evaluations and local school systems. Considering the impact that teachers have on students' success, critiques that locally based evaluation systems may make removing "bad" teachers who are protected by tenure due to lack of evidence of their ineffectiveness and that personal preferences of an administrator may keep ineffective teachers in the classroom, are valid (Chetty, Friedman, & Rockoff, 2011; Hanushek & Rivkin, 2010; Haycock, 1998; Nye, Konstantopoulos, & Hedges, 2004; Rowan, Correnti, & Miller, 2006; Sanders & Rivers, 1996; Schacter & Thum, 2005). Over time, such critiques have led to formal policy changes affecting the ways in which teachers are evaluated.

This chapter briefly delineates the history of teacher evaluation in the United States with a focus on the last two decades, which highlight a marked shift from local control over evaluation systems to the use of various interventions from both state and federal governments. The second half of the chapter describes the educational context of the study state, North Carolina.

A Brief History of Teacher Evaluation

Historically, decisions about hiring, evaluating, retaining, and firing teachers have been made at the school-level and school administrators have generally been able to exercise a great deal of freedom in selecting teachers for open positions (e.g. Tyack & Cuban, 1995). For instance, the evaluation of schools in general can be traced back to the Common School Era in Massachusetts where Horace Mann rode from school to school writing analyses of each location he visited (Mann, 1868). And while evaluation policies have not always been formalized, teachers have always been held accountable to someone for something, whether it be the tidiness of the classroom or whether students could recite memorized text to an audience. What teachers were held accountable for and who they were accountable to has varied, but such accountability was always tied to the retention of a teaching position.

However, in the early days of American schooling, the values that defined good teaching were determined and defined at the local level. Over time state, and later, federal government became increasingly involved in matters of school regulation, including the regulation of teacher quality, which includes evaluation. While full federal intervention in public schools is fairly recent, the first attempts to federally influence education can be traced to the aftermath of the Civil War with Congressional debate over establishing a federal Department of Education in 1866, followed by the failed Hoar Bill of 1870 which attempted to establish federal takeover of public schools which were failing (Newman, 2013). Such early attempts failed to be implemented, but presented policy frameworks which manifest in contemporary federal education policy.

In the last two decades, as federal influence has increased via directing state-level policy in schools through both mandates and incentives, some traditionally locally held powers, such as control of teacher evaluation, have shifted and become more centralized, at least in part, at the state level. This shift has been gradual, with early critics pointing out that such changes have increasingly de-professionalized teaching. For instance, Giroux (1985) contended over 30 years ago that curriculum policies disempowered teachers and reduced their status to that of a high-level technician of objectives and goals created by people with no experience with classroom realities. Although this shift in power structure has occurred gradually over time, the 2009 Race to the Top (RttT) initiative incentivized states to create legislation that sometimes drastically changed local districts' and schools' ability to control how they recruit, compensate, and maintain their teaching workforce. In this chapter, details about these other policy points may be included in cases where such policies are linked. Finally, I briefly describe how teacher evaluation policy has evolved to its current nature at the time of this study.

As the U.S. has undergone a shift from a tradition of local control to a more centralized governance structure, the values defining a "good" education and "good" teaching have also shifted and become more universally defined by policy. What is valued in education is not something necessarily stated explicitly in most policies, but instead values are something that can be decoded from various sources such as student learning standards, classroom curriculum, teacher preparation requirements, professional development components, professional licensing requirements, and performance evaluations. Evaluations and other accountability mechanisms may be the most influential component of defining educational values because such measures explicitly state what should be accomplished in the classroom and to what degree it should be accomplished. Likewise, the shift in governance has been accompanied by an increase in accountability from local to external (non-local) sources, which has created a greater and perhaps more narrowed consensus of what is valued in schools. Additionally, it is unclear how such high-stakes accountability policies derived from state governance affect the work of teachers. It is possible that such policies, when centralized at the state level, may bear more influence on individuals than previous local evaluation policies and therefore, cause greater impact to the teaching workforce.

Although centralization is broadly defined as the consolidation of power at a higher level of government, at issue here is the transfer of power over decisions regarding the teaching workforce from local governing bodies to the state level, which often occurred under the direction and guidance of the federal government. While some aspects of this move to centralization at the state level, such as the creation of state teacher certification (the first state tests emerged in the 1860's which were followed by university preparation programs in the early 20th century), occurred much earlier, much of the more recent evidence of this shift can be seen in what has been termed the "Accountability Movement" (Vinovskis, 2009). The Accountability Movement included a shift to standards-based reform and outcome-based education models.

Mintrop and Sunderman (2013) describe the evaluation movement that has accompanied increased centralization in school governance as occurring in three waves. These waves offer a framework for understanding the progression of school accountability policy and indicate that over the last two decades student test scores on standardized tests have increasingly served as a proxy for student learning. Additionally, states or localities have used these measures to influence teacher pay, retention, or promotion. The first wave of this accountability involved experiments in states, such as Texas, and localities, such as Chicago (Mintrop & Sunderman, 2013). The seeming success of these smaller scale experiments largely inspired the second wave of reform.

The second wave formed at the national level with a series of educational goals first presented by President George H.W. Bush and his America 2000 plan, and later refined by President William Clinton's Goals 2000: Educate America Act (Vinovskis, 2009). Both plans introduced national goals for education and were precedents for President George W. Bush's No Child Left Behind (NCLB), a renewal of the Elementary and Secondary Education Act of 1965 (ESEA), which was a law passed by President Lyndon Johnson that established federal funding for schools as part of his "War on Poverty." NCLB introduced federal guidelines for states as well as punitive measures for schools failing to meet expectations (Vinovskis, 2009). Additionally, after the passage of NCLB in 2001, test scores became a main component of measuring the effectiveness of individual schools and districts, representing the second wave of accountability, where failure to make targeted improvements in different measures led to sanctions including the possibility of state takeover (Mintrop & Sunderman, 2013). The second wave marked an era of sanctions where federal guidelines required state takeovers or closures of schools deemed to be "failing" to make established growth guidelines. These takeovers differentially impacted poor socio-economic areas and occurred primarily in urban districts, such as Chicago.

America 2000, Goals 2000, and NCLB paved the way for the Obama administration's Race to the Top (RttT) Initiative of 2009, which was followed shortly by the NCLB/ESEA waiver program, which prompted states to undergo several legislative changes to reform education in order to compete for money to supplement state budgets, or in the case of the waivers, to seek relief from NCLB mandates. The RttT initiative was funded by the American Recovery and Reinvestment Act of 2009 (ARRA) which allocated $4.35 billion dollars for the RttT program. Although only 12 states received the funds, the application process required changes to existing school systems and governance structures at the state level and legislative changes occurred in all applying states. One required change for states applying for RttT funds was to alter teacher evaluation policies and adopt student growth as a main measure of teacher evaluation as well as standardizing previously used observation protocols (US Department of Education, 2009). Thus, through RttT federal values have influenced the states' assumption of previously-held local powers over the teaching workforce.

So, along with other changes to school policy, RttT enticed states to implement new personnel laws, including revamping teacher evaluation systems to include student growth measured by state test scores along with the use of standardized observation data as part of a requirement for multiple measures of evaluation (US Department of Education, 2009). Furthermore, these evaluations were required to be attached to personnel retention decisions, which prompted states to make changes that eliminated or reduced tenure. In most cases, these personnel laws were changed along with laws that permitted greater numbers of charter schools, increased alternative pathways into the teaching profession, mandated the creation of statewide data systems to serve as repositories of information on both students and personnel, and made changes to state-level student academic standards, largely through the adoption of the Common Core State Standards (CCSS).

Thus, in many states, the RttT legislation greatly impacted the way school was managed including the ways in which teachers were hired, retained, and fired. It is important to note that due to the simultaneous adoption of multiple policies, policy actors (including teachers) may be unable to discern these as separate, distinct changes. In other words, changes to things like evaluation, tenure, and teaching standards, having occurred concurrently may appear like a "package deal" to teachers who are influenced by all components of the package simultaneously.

Additionally, states, partially based on resource disparity and partially based on existing systems and traditions, have varied greatly in their approaches to meeting these new federally-inspired laws.

Under NCLB, schools faced sanctions for failing to grow student scores in accordance with goals set for Annual Yearly Progress (AYP). So, the shift to using student test scores as a proxy for teacher rather than school effectiveness represents the latest incarnation of test scores as a proxy of student learning and represents the third wave of accountability as espoused by Mintrop and Sunderman (2013): one that is focused on the effects of the individual teacher. The legislative changes in state level teacher evaluation policy that occurred during RttT coincide with the third wave.

An unintended consequence of these legislative changes was a further narrowing of what policy values as important in education as tested schools undergo more intense microscopic examination under these teacher-focused policies. However, state policies attempt to mitigate this by pairing the student effectiveness component of evaluations with standardized observations to create multiple forms of measurement. Evaluations using both observation and student growth measures are intended to be more concrete and uniform across systems than precursors which were often designed at the local level based on local values and priorities. The rationale behind the change was that multiple measures of teacher effectiveness will produce a fairer rating and better feedback than if districts relied upon a single measure instrument. However, it is important to remember that despite these federally inspired changes, there are still policy discrepancies across and even within states.

Furthermore, the publicity accompanying such legislative changes often touted teacher evaluation policy as a much needed and previously unexplored area of educational governance, which often obscured the fact that teachers have always been held accountable for their practice in some way. What has changed is the technology behind teacher evaluation, the shift in educational values linked to such measures of teacher quality, and the demand for new infrastructure required by utilizing sophisticated psychometric techniques such as VAMs (Anagnostopoulos, Rutledge, & Jacobsen, 2013b). Such infrastructure has not previously been evident in U.S. schools which have lacked a system of common evaluations, standards, and frameworks; this makes teaching in Americans schools much different from other skilled, service occupations (Cohen, 2011). As such, many have described the shift from NCLB to the RttT requirements for teachers to be a shift from a designation of "highly qualified" to one of being "highly effective" as localities are asked to focus less on what qualifications teachers bring to the job, but rather what sorts of results are produced by teachers (Powell, 2013).

NCLB remained in effect until December 2015. In 2011, shortly following the announcement of the RttT competition, then U.S. Secretary of Education Arne Duncan instituted a waiver program whereby states could seek flexibility from specific provisions of the federal legislation, most specifically the unobtainable 100% proficiency requirement. As of 2014, 42 states and the District of Columbia had applied for and obtained ESEA waivers, but many lawmakers viewed them as an unconstitutional subversion of federal policy in exchange for the adoption of executive branch preferred policies (Epenbach, 2014; Umpstead & Kirby, 2012). Regardless, sweeping legislative changes occurred in many states due to a combination of RttT application and NCLB waiver requirements.

The third generation that Mintrop and Sunderman (2013) described is the current wave at the time of this dissertation and includes the latest federal influence, the RttT competition and its inspired legislation. What distinguishes this third wave is increased focus on the accountability of individuals rather than entire schools or systems. In evaluation, this has manifested as evaluation systems that include psychometric measures meant to gauge an individual teacher's exact effect on a student as measured on a standardized test. The third wave has also brought about a standardization of observation protocols for teachers and a greater value placed on teacher performance on evaluations when considering job retention.

It is notable that most states that changed their laws did not receive the RttT funding; however, most did eventually receive a waiver from NCLB compliance. Therefore, most states were tasked with implementing unfunded, mandated changes to schools and systems. What is of interest here are the changes related to the standardization of teacher evaluation and the narrowing of accountability focus to the level of the individual. Thorn and Harris (2013) characterized this shift as follows: "This shift in the way we measure success in education represents a sea change, with consequences for the way schools operate as well as for the individual autonomy that teachers came to expect during the past half-century (p. 57)," a sentiment that suggests that macro-level policies can and do effect teachers at the classroom level.

A Brief Legal Review of the Governance Shift as Related to Evaluation

Several law reviews acknowledge how both NCLB and RttT have affected educational governance structures at the state and local level. For instance, Garda and Doty (2013) argued that NCLB compelled states to implement "far ranging governance reforms for failing school districts and Title 1 schools," but that these efforts at the individual school-level have failed (p. 2). RttT, however, incited governance changes at the state level. The review outlines the requirements of NCLB's annual yearly progress (AYP) requirement and discusses many of the legal issues that resulted from such mandates. For instance, *Reading School District v.*

Department of Education illustrates one of many failed attempts of schools and districts to contest the labeling of schools as not meeting AYP (Garda & Doty, 2013). While NCLB was only enforceable through the mechanism of withholding federal Title 1 funds, RttT enticed states to change laws to meet federal values and priorities through grant applications and NCLB waivers. Garda and Doty further pointed out that the failures of NCLB and RttT to create meaningful reform have not been a result of complex legal issues or lawsuits, but rather from political resistance (2013). This suggests that the issue states have with federal influence is a result of changes to the power structure and governance.

Umpstead and Kirby (2012) also acknowledged several of the high-profile lawsuits that challenged NCLB, particularly those focused on the limited funding available to states who were tasked with implementing what was essentially an unfunded federal mandate, such as: *School District of Pontiac v. Secretary of the Education Department* and *Connecticut v. Duncan* as well as those regarding NCLB's effects on student achievement, such as: *Levi v. O'Connell, Board of Education of Ottawa Township High School District 140 v. US Department of Education, and Coachella Valley Unified School District v. California*. This piece noted that the NCLB waivers may have been unconstitutional due to coercing states to adopt other policies found preferable by the Obama administration, and initial drafts of the Obama administration's ESEA reauthorization included many of the same provisions present in the RttT and NCLB waiver applications, which led to a delay in the law's reauthorization (Umpstead & Kirby, 2012). Issues of teacher quality were also addressed, most specifically through NCLB's highly-qualified teacher provision, which created a variety of designations across states trying to meet the mandate. For instance, the lawsuit *Renee v. Spellings* challenged California's designation of teachers without full certification as highly qualified, an opinion that was upheld in the appeal *Renee v. Duncan*, leading to Congress responding by adjusting the law and further illustrating the complex relationship between law, governance, and education (Umpstead & Kirby, 2012).

Furthermore, Barnes (2011) outlined the history of ESEA leading to RttT and contended that given the results of previous federal initiatives, the only benefactor of resulting RttT legislation was “big government” and contended that the program led to the violation of individual liberties. Her arguments are linked mainly to previous litigation that resulted from attempts to create standards in education, yet the criticism that RttT violates individual liberties could also be applied to teaching issues, such as the loss of due process rights through discontinuing tenure and the loss of a fair and transparent evaluation procedure.

Similarly, Powell (2013) directly investigated issues of teacher quality including the weakening of the tenure system. She contended that tenure is not the reason that ineffective teachers become difficult to fire, but rather that this is due to the ineffective and unreliable procedures utilized in teacher evaluation. Citing studies such as the New Teacher Project’s “Widget Effect” (Weisberg et al., 2009), Powell stressed that states need to not only adopt legislation required to change evaluation procedures, but also to implement strategies to attract and retain effective teachers; in her view, this includes a streamlined evaluation process and the maintenance of due process rights.

Teacher Evaluation in North Carolina

North Carolina was an ideal location for examining the convergence of state-level evaluation policy and classroom conditions due to its strong, pre-existing statewide evaluation policy. Unlike many other states, North Carolina designed a precise evaluation instrument that all districts were required to utilize that pre-dated RttT (Table 1). This existing system was one reason why North Carolina was able to score highly on the RttT application and become one of the states that received funding through the program. Upon the announcement of the RttT competition, North Carolina broadened the evaluation to include a value-added model (VAM) of student performance and changed infrastructure related to the evaluation to accommodate the new policy. North Carolina also was one of the 12 states that received RttT funding in 2010. Because the statewide evaluation system has been in place in some form prior to RttT and has been ingrained as part of teacher practice for many years, North Carolina schools are an excellent place to examine how such policies impact classroom practices.

Table 1

Timeline of Educator Evaluation Changes in North Carolina Since 2009

Year	Relevant Legislation	What happened
Dec 2009, Updated Feb 2015	TCP-C-004 16 NCAC 06C .0503	Establishes three types of evaluation cycles and a process for performance appraisal.
Dec 2009	TCP-C-019	Teacher and principal evaluations must be submitted to the state superintendent annually.
July 2011	115C-333	State must be notified of employee dismissals.
Aug 2011	TCP-C-022	All systems must evaluate all teachers annually and must include the student growth component.
August 2012	TCP-C-006	Standard six, the student growth standard, is added to the evaluation.
August 2013	Current Operations and Capital Improvements Appropriations Act of 2013, ch. 360, 2013 N.C. Sess. Laws 995	One-year contract structure is initiated for teachers who have not met career status recognition, requiring full annual evaluation cycles for all teachers without career status indefinitely. Permanent elimination of all career status designations to occur in 2018 (currently ruled unconstitutional).

In 2009, the North Carolina General Assembly passed a mandate to create teacher evaluation procedures that supplemented and supported newly-created State Board of Education requirements under TCP-C-006 (North Carolina State Board of Education, 2012b). The policy also specified a process for professional growth plans for teachers. Meanwhile, TCP-C-019, which was created in December 2009, specified that all teacher and principal evaluations must be submitted to the state superintendent annually (North Carolina State Board of Education, 2012a).

By the time TCP-C-006 and TCT-C-019 had been passed, the state had already begun a massive state-wide roll out of what was then termed the "New Teacher Evaluation." The evaluation at this time consisted of five observation standards and included pre- and post-conferences as well as a year-end summative conference. Training was provided for administrators to ensure fidelity to the instrument and training was also provided to teachers. These trainings occurred at the school-level, were provided by staff from the North Carolina Department of Education, and were mandatory for all teachers. Upon applying for RttT funding, North Carolina added a sixth standard which accounted for "student growth." Trainings on this standard occurred in spring 2012 and the standard was included with 2012-2013 evaluation onward (North Carolina State Board of Education, 2012a). In 2016, the North Carolina Department of Education announced that student growth will be removed as a stand-alone standard and would instead be incorporated into the other five standards. However, the logistics of that transition were not yet clear at the time of this study.

So, under the system which was current at the time of this study, all teachers in the state were measured against the state instrument consisting of five observation type standards and one student growth standard. State Board Policies and Statutes TCP-C-004, most recently updated in February 2015, established the performance appraisal process including: the creation of three types of evaluation cycles dependent on a teacher's certification and administrator assignment, and a process including training, orientation, self-assessment, observation, pre- and post-conferencing, and summative evaluation. At the time of this study, the evaluation is administered differently depending on whether a teacher has received "career status" in their district. A one-year contract structure was instituted in 2013 under the Operations and Capital Improvements Appropriations Act of 2013, which required teachers who had not achieved career status by that time to undergo a full evaluation cycle each year indefinitely (four formal observations). In other words, teachers who did not earn career status prior to 2013 are no longer eligible for that designation and the evaluation process follows that distinction. The law also stated that career status would be removed from all North Carolina teachers at the conclusion of the 2017-2018 school year. Court litigation and several rounds of appeals followed the passage of this act with the most recent update being that the one-year contracts for those who never made career status has been upheld, but the repeal of tenure for those with career status had been unanimously deemed unconstitutional by the NC Supreme Court in June 2015. However, the law remains active at the time of this study.

Therefore, a teacher who has career status would be required to complete only an abbreviated evaluation each year consisting of two abbreviated observations that may not cover all the standards. The exception is teachers who are renewing their licensure in the current year who are also subject to a more intense evaluation cycle consisting of four observations, regardless of having career status. These requirements can be modified based on administrator discretion and teachers may receive more evaluations than what the state requires if administration decides. As previously stated, a repeal of career status entirely would mean that all teachers in the state would have to undergo a full evaluation cycle annually.

Moreover, the results of each teacher's individual evaluation are reported and tracked at the state level. Therefore, state-level administration can gauge the effectiveness of any teacher in the state, according to the evaluation instrument, at any time using the state data system. A teacher's effectiveness is tracked throughout their career so long as they remained in the state of North Carolina. As a growth instrument, teachers are expected to "grow" on their evaluation throughout their career. This is a marked departure from previous systems where teachers could leave past effectiveness ratings behind by obtaining a new job in a different school system. Furthermore, 115C-333, also passed by the North Carolina General Assembly, requires the notification of the State Board of Education upon dismissal of employees. These policies demonstrate the power over the teaching workforce that is held by state-level institutions following RttT. The longitudinal tracking of teachers at the state level coupled with the legislation attaching evaluation to employment retention makes this evaluation policy particularly high-stakes for teachers in North Carolina. Additionally, because the policy and the instruments were designed at the state level, and because evaluators' scores are tracked by the state, it is possible that what the state values in education overtakes what local administration values about quality teachers and teaching.

North Carolina is an ideal site for research on the relationship between evaluation and practice because the state-level policy is so strong. Not only are all teachers subject to the same evaluation protocol, but the results are reported directly to the state-level. Additionally, large numbers of teachers do not have career status and as such are under one-year contracts and subject to full evaluation cycles consisting of at least four observations annually. Moreover, the state has gone to great lengths to eliminate career status altogether, which would make all teachers subject to one-year contracts and full evaluation cycles if the Supreme Court decision is not upheld. These changes in career status, coupled with evaluation results being a top consideration for lay-offs in cases of reduction in force, make the evaluation policy high-stakes for teachers in North Carolina.

CHAPTER 3: Literature Review

In this chapter, I review the literature relevant to my research questions. In doing this, I build a theory upon which my research is based. I first approach the two aforementioned evaluation policy assumptions, that evaluation simultaneously motivates teachers and provides feedback to improve practice, by separately exploring the ideas behind teacher motivation and the relationship between feedback and teacher practice. I primarily draw on two bodies of literature to further situate my study. Because there is a gap in the literature regarding the relationship between teacher evaluation and teacher practice, I first review literature on how teachers have responded to other external accountability pressures, specifically pressures resulting from NCLB. Secondly, I review literature on how teachers modify practice to accommodate curriculum reforms. Additionally, to better understand the differences in school contexts, I review literature on the relationship between school leadership and evaluation implementation.

Examining Policy Assumptions

Assumption 1: Teacher Motivation

This sub-section addresses the policy assumption that evaluations are necessary because teachers need to be rated, sanctioned, or rewarded in order to be motivated to do a better job. Firestone (2014) identified two theories of motivation that guide thinking about evaluation. The first theory Firestone describes is an economics-based theory focused on external rewards and the second theory is based in psychology and focused on intrinsic reward with teachers improving practice through assessment, feedback, training, and professional development.

Extrinsic motivation. The theory that teachers are most motivated by external forces comes from the field of economics. Such thinking is reflected in a number of existing financial policies such as career ladder pay scales, bonuses or salary increases for passing proficiency exams, recruitment and retention bonuses, and performance or merit-based pay. At issue in the policy context of this study is what Firestone argues is “the most powerful incentive… access to employment itself” (2014, p. 102). In North Carolina, teacher evaluation is the top criterion for deciding which teachers will be removed from employment when there is a reduction in force in a school system. Additionally, evaluation results are reported to the state and past performance is accessible to other potential employing schools statewide. In contrast to teachers, students under the same system “have no direct incentives to perform in such schemes, apart from whatever pressure their teachers can create” (Cohen, 2011, p. 74). This means that teachers often must persuade students that academic work, specifically the test upon which part of teacher observation scores are based, is even worth doing (Cohen, 2011). As a result, teachers may alter behaviors to try and improve student achievement in ways that would favorably influence results.

Aside from the idea that poorly-performing teachers should be removed from the system, extrinsic factors can also lead to teachers self-selecting out of the system. For instance, research shows that teachers leave schools when they do not receive competitive salaries and that qualified individuals may seek employment in other sectors (Ingersoll & May, 2012; Johnson & Birkland, 2003). Because North Carolina offers a statewide salary schedule with limited local supplements, there is little financial competition between districts and teachers may move to nearby states or remove themselves from education careers altogether. Alternatively, extrinsic theory also means that teachers may choose to continue in the profession even if teaching is not their main priority. For instance, studies in American education suggest that students attach little importance to academic learning over practical knowledge, and may hold their teachers in little esteem (Cusick, 1983; Powell, Farrar, Cohen, 1985). As a result, some teachers focus on other aspects of school, such as coaching or relating to students, often as part of the negotiation process to make their jobs bearable (Cusick, 1983). Therefore, teachers who are driven by external motivation may stay in the profession for an income, even if teaching is not an individual priority.

Intrinsic motivation. Intrinsic motivation theory stems from the belief that people are rewarded by the feedback they receive from their work, and that they feel good when they are performing well (Deci & Ryan, 1996; Hackman & Oldham, 1980.) In the simplest terms, this means that someone who is intrinsically motivated feels good when they do well. Firestone (2014) argued that in general, those who are motivated internally experience both autonomy and self-efficacy and therefore evaluation should create rewards and contribute to the creation of rewarding conditions. In a previous review of working condition studies, Firestone and Pennell (1993) found that 10 out of 13 studies confirmed this relationship between teacher autonomy and teacher commitment, a condition that they contend is similar to motivation. Similarly, in an earlier critique on curriculum policy, Giroux argued that a technocratic approach to policy is grounded in the assumption that teacher behavior needs to be controlled and made consistent and predictable across all contexts, thereby reducing teacher autonomy to plan and develop curriculum and instruction to instead teach to a test (1985). Firestone (2014) also contended that research on self-efficacy (e.g., Bandura, 1997) and teacher efficacy (e.g., Tschannen-Moran, Woolfolk Hoy, & Hoy, 1998) suggests that competence and expectancy are motivating forces for teachers. In this case, competency means that the individual has the capacity to carry out the expected tasks and expectancy implies that the actions of the individual will lead to an intended outcome.

Teacher competency and expectancy may also vary based on several classroom level conditions that teachers may be unable to control, such as teaching assignment (Ball & Bass, 2000) and student interaction with classroom materials (Cohen, Raudenbusch, & Ball, 2003). However, more systemic conditions such as administrative support, adequate physical facilities, adequate instructional materials, and realistic workloads also may influence a teacher's competency and expectancy (Firestone & Pennell, 1993). Additionally, research suggests that teachers are more motivated in schools that are orderly, have adequate school discipline, and are not overly punitive (Firestone & Rosenblum, 1988; Garet, Porter, Desimone, Birman, & Yoon, 2001; Ingersoll & May, 2012; Johnson & Birkeland, 2003; Kushman, 1992). Firestone contends, "The opposite of the fully autonomous individual is the person performing an activity under duress" (2014, p. 101). Additionally, the importance of evaluation conditions is evident in Cohen's (2011) description of how the work of teachers is regulated by the society, economy, and culture around them and that a lack of consensus about educational results can increase uncertainty and dispute in a school whereas such conditions may not exist in a more cohesive school with individuals of similar ability.

Assumption 2: Feedback

Aside from ineffectively rating teachers, criticism has also abounded that previous evaluation systems did not provide enough information to improve teacher quality through feedback. Boyd, Grossman, Lankford, Loeb, & Wyckoff (2006) found that without useful feedback, most teachers' performance plateaus by their third or fourth year on the job. Yet, locally developed evaluations used in the past have often been criticized as providing only a cursory review of teaching practice. Furthermore, research suggests that feedback that directly stems from the work itself can contribute to enhancing teacher competence and intrinsic rewards

(Hackman & Oldham, 1980). Most new teacher evaluation systems, including the one examined in this study, use multiple measures in combination to evaluate teachers. A typical manifestation is a combination between a standardized observation protocol and a value-added measure of teacher effects based on student standardized test scores, which is what is utilized in North Carolina at the time of this dissertation.

One justification of using a system of multiple measures is that it theoretically will yield multiple types of feedback for teachers to use to improve practice. Additionally, the standardizations will define focal points deemed important. And while feedback has historically come directly from students (Black & William, 2009; Hart, & Murphy, 1990), formal teacher evaluation could provide feedback through both quantitative measures of student achievement and structured observation tools that are now part and parcel of teacher evaluation policy.

Despite current policy often mandating the use of multiple measures, classroom observations are often viewed as the instrument that is mostly like to provide actionable guidance on how to improve teaching. This is because unlike the summative assessment produced with student achievement data, observation protocols are often accompanied by post-conference reflection between the observer and the observed. Additionally, there is some evidence that when teachers are provided scores and feedback from standardized protocols by a research project staff member or an administrator, respectively, they improve their practice (Allen, Pianta, Gregory, Mikami, & Lun, 2011; Taylor & Tyler, 2011). Feedback, however, can differ greatly depending on the person who is providing it. For instance, successful learning can have varying definitions from individual to individual (Cohen, 2011). So, it is possible that the quality of feedback a teacher receives will be influenced by an evaluator’s values despite the standardization of observation protocols.

Additionally, there is some emerging evidence that what an observer chooses to emphasize for improvement may be determined by the subject area being observed. For instance, Bell et al. (2015) found differences in the rank ordering of teachers when different protocols (general versus subject specific) were used. Additionally, rank ordering differed based on the subject area taught by the teacher compared to the observer's subject area background. This study also found that note-taking and feedback patterns from evaluators differed depending on the subject matter background of the observer and whether there was alignment between an observer's background and the subject being taught. The differences were more pronounced in mathematics, which suggest significant complexity in the ways that protocol, subject matter, and observer background intersect.

Similarly, evidence exists of such differences in literature on how potential observers deal with different types of reform. For instance, in a study of 15 elementary school administrators and 15 curriculum coordinators, Burch and Spillane (2003) found that more emphasis was placed on teacher inputs and building literacy across subject areas with literacy reforms while math reforms focuses on sequenced instruction and external supports. Therefore, the quality of the feedback received may be dependent on many factors including the subject being taught and the background of the individual providing it. Both are likely to play a role in whether an individual teacher finds observation feedback useful.

Teacher Responses to External Accountability Pressure

While there is a gap in research regarding the relationship between evaluation and teacher classroom practice, research on other external accountability policies based on student testing results is extensive and has demonstrated unintended effects on the teaching workforce, primarily in the form of turnover, as well as on practices in the school or classroom. For instance, “gaming” refers to engaging in strategic behaviors that will increase reported performance without making gains in actual student performance. Attempts at gaming can range from outright cheating and changing answers (Jacob & Levitt, 2003) to more benign techniques such as changing the quality of student lunches during testing (Figlio & Winicki, 2005) or moving the teachers with the best records of producing gains to tested areas (Cohen-Vogel, 2011; Grissom, Kalogrides, & Loeb, 2012). I briefly describe some commonly referenced issues with external accountability: teacher turnover, narrowing of curriculum, prioritizing the teaching of strategies over curriculum, and the triaging of students.

Turnover

Research suggests that external accountability pressures impact the teaching force, particularly in high-need schools. For instance, Clotfelter et al. (2004) suggested that low-performing schools at risk of performance sanctions experienced negative effects on retention rates and on the probability of filling a vacancy with a high-quality teacher. If such evidence is true for sanctions at the school-level, then it would be reasonable to assume that these negative effects could persist, and possibly be amplified, when sanctions are applied at the teacher level through value-added models (VAMs) and standardized observation protocols. Also, dismissing teachers based on poor student test growth becomes problematic when dealing with low-performing schools that may already be experiencing staffing difficulties. In such cases, it becomes clear that dismissing poor-performing teachers based on evaluations does not offer the sole solution to the issue of consistently low-performing schools.

Research suggests that teacher turnover for any reason comes at great financial cost to schools and educational costs to students (Ingersoll, 2001). While changes in evaluation were largely driven by criticisms of locally based observation, most of the practical problems identified with current evaluation policy focus on the use of student growth driven VAMS. This creates a conundrum where the costs of losing an effective, but misidentified teacher must be weighed against the costs of leaving students with an ineffective teacher who would not be dismissed without the use of VAMs. Policy makers should also consider the costs of possibly keeping a bad teacher who was misidentified as effective and may be difficult to dismiss. Raudenbush and Jean (2012) argue, "Falsely identifying teachers as being below a threshold poses a risk to teachers, but failing to identify teachers who are truly ineffective poses risks to students" (p. 2). Numerous researchers report that the risk of misidentification of teachers is high and widely variable depending on the model and confidence interval used (Raudenbush & Jean, 2012; Goldhaber et al., 2013).

Similarly, Goldhaber et al. (2013) demonstrated how, depending on model specifications, teachers could easily switch the quintile in which they are assigned, showing that the most reliable use of VAMs can be found in separating only the truly outstanding teachers from the truly terrible, something that may likely be already known in a school, and that the middle quintiles show extreme variation based on the specifications used. To this extent, VAMS are prone to the same criticism of previously used local level evaluations when teachers are not being accurately labeled. As Harris (2009) points out, this unreliability in VAMs provides little in terms of formative feedback about a teacher's practice and instead serves to summatively signal quality, something that could be dangerous given the extreme variability described above when attached to high-stakes policies. Again, aside from potential financial consequences, such systems are likely to also challenge teachers' feelings of competence and efficacy.

Additionally, the potential inequity and instability of the VAM instrument may pressure certain teachers to exit the system. Although VAMs have been adopted in many states, including

North Carolina, this adoption has been highly criticized by both scholars and practitioners. One issue is the lack of tests for all grades and subject levels. Under NCLB, states had to create tests in some, but not all, grades and subjects. As some researchers have suggested, this lack of tests is most alarming at the high school-level, where NCLB mandates only one test in each subject area even though each student has different teachers for each of several subjects each year (Goldhaber et al. 2013; Harris, 2009). Harris (2009) also raises the question of how VAMs will be able to account for the possible effects of other teachers (particularly at the high school-level where a student is enrolled with several instructors simultaneously), teamwork among staff, and peer effects.

Furthermore, the aforementioned challenges in the calculation of VAMs suggest that, at least at the high school-level where there are more specialized courses, there may be an additional challenge in shifting to a VAM that assumes that student ability is comparable across all subject areas (Goldhaber et al., 2013). Many VAM models use prior test scores to predict future achievement, which is problematic in more specialized courses and curricula, such as Physics, which would not have a prior test. This disconnect challenges teachers' feelings of efficacy and competence, which may in turn drive teachers of more specialized subjects from the workforce. Regardless, replacing any teacher comes at a financial cost and instability in a school's workforce can carry educational effects as well, regardless of whether the teacher is removed or leaves and regardless of whether that teacher was effective (Ingersoll, 2001). Therefore, it is important to consider how teacher evaluation policies may be linked to teacher turnover and retention.

Aside from the threat of job loss when accountability is attached to high-stakes personnel decisions, the research on turnover is important to consider when thinking about the relationship between motivation and evaluation. If evaluation elicits feelings of incompetence in an individual it may affect their intrinsic motivation. Teachers may choose to leave a school in favor of another school or job that provides the types of intrinsic rewards necessary to foster work satisfaction. Likewise, if evaluation is attached to extrinsic rewards such as bonuses or job security, then individuals may choose to leave the system in favor of positions that are more extrinsically rewarding and financially secure.

Narrowing Curriculum and Teaching Testing Strategies

While the knowledge and skills of a teacher are important, the work of teachers is also entirely dependent on the willingness of students to participate in learning. As such, the negotiation of curriculum is a key event in classrooms. Cohen (2011) argues that, "practitioners must supplement their expertise with client's consent and with the knowledge and skills that clients bring to bear" (p. 12). As such, teachers often find ways to anticipate what students will find interesting in order to negotiate content and workload (Cohen, 2011; Powell et al., 1985). Accountability has added an extra layer to this dilemma as teachers may now feel pressure to emphasize certain areas of the curriculum known to be emphasized in assessments. For instance, some research suggests that an increased focus on testing outcomes in certain subjects has resulted in a narrowing of curriculum that increases as external pressures increase (Carnoy & Loeb, 2002; Ladd & Zelli, 2002; Rothstein & Mathis, 2013). Such work follows the logic of Milgrom and Roberts (1992) and the principal agent theorem, which contends that in organizations with multiple goals, agents will focus on rewarded goals at the expense of other goals.

Additionally, American students are tested more than any other students in the world, yet there is little agreement over what should be in tests and there is often considerable variability among the curriculum standards and tests of the same subject (Conley, et al. 2011; Floden, Porter, Schmidt, & Freeman, 1980; Porter, McMaken, Hwang, & Yang, 2011; Porter, Polikoff, & Smithson 2009). Cohen (2011) describes how this lack of agreement can lead to some teachers aligning content with standardized tests whereas others may select from a textbook or a workshop, or simply choose to teach what they learned as students. Therefore, standardized testing has done little to create uniform curriculums across locales and may actually result in the narrowing of curriculum to meet specific demands of specific evaluations.

Also, teachers may forego teaching curriculum altogether and devote lessons to teaching test-taking strategy rather than content. For instance, research suggests that VAMs potentially reward teachers who use a curriculum focused on testing or testing strategy rather than actual subject matter (Carnoy & Loeb, 2002; Goldhaber, et al., 2013; Ladd & Zelli, 2002; Mintrop & Sunderman, 2013; Rothstein & Mathis, 2013). Therefore, teachers may feel pressure to devote class time to teaching testing skills rather than actual components of the subject area.

Concerns about narrowed curriculum or replacing curriculum with teaching test strategies are important to consider as the VAMs tested in the MET study are prone to large error with a correlation of around 0.5 for elementary teachers, and that this error would increase as teachers focus more on the goal of increasing scores and avoiding sanctions (Rothstein & Mathis, 2013). In other words, the greater the risk of sanctions attached to scores, the greater the risk of focus on the tested curriculum and testing techniques at the expense of other areas of curriculum. Therefore, it is important to consider how evaluation may influence what is taught in a classroom. It is possible that teachers may be adapting the curriculum they teach to address components that are more likely to impact their evaluation. This could be true in regard to narrowing the curriculum, but it is also possible that teachers may select certain lessons that they feel will be more appealing for instances when they know they may be formally observed. It is also possible that teachers may modify how they teach, such as by employing the use of more assessments that look like those formally used for evaluation, or picking teaching strategies that may be more appealing to observers, such as employing technology the day of the observation or utilizing a particular method, such as Socratic seminar, if it is thought an observer may score more favorably.

Triaging Students

Research has determined that another popular method of gaming in schools involves removing low-performing students from the test pool. This can be done in a variety of ways. For instance, a study by Figlio and Getzer (2002) showed that students who were low-income or previously low-achieving in six large Florida districts had been categorized as students with disabilities (SWDs), a category that was exempt from testing at the time of the reassignments, at a rate much higher than prior to the implementation of accountability policy. Similarly, a study of over 41,000 disciplinary events in Florida schools suggests that schools assigned substantially harsher punishments to low-achieving versus high-achieving students with a significantly increased gap during the testing period (Figlio, 2006). Such practices served the purpose of removing the scores of students who may be poor achievers.

Although most of the available current studies extend to school-level accountability policies and school-level gaming practices, there is evidence to suggest that similar actions may also occur at the classroom level. For instance, Booher-Jennings (2005) described how a school in Texas participated in "educational triaging." Under this system, resources were diverted towards students who were predicted to be at threshold levels of passing the state assessment as well as towards students who were counted towards the school's overall accountability rating.

Similar behaviors were observed in a study in Chicago where teachers diverted more attention to students near the pass threshold (Neal & Schanzenbach, 2010). It is likely that teachers will continue to engage in similar behaviors with new accountability policies focused on the level of the individual teacher. Therefore, it should be considered that teachers may direct focus on certain students based on their evaluations.

Teacher Responses to Curriculum Reform

Some of the assumptions behind teacher evaluation policy are based in economic theories. Specifically, these assumptions originate from the idea that teachers will behave as rational actors within a system, and that given increased pressure, teachers will perform better (Milgrom & Roberts, 1992). However, such economic views fail to account for the manner in which evaluation reform is embedded in existing institutional structures. So, while economic theories inform the construction of the policy, such theories are unable to predict the behavior of the actors affected by such policy. While there is a gap in the research about the ways in which teachers may respond to evaluation reform, there is a lot of information available on how teachers respond to curriculum reforms. The research on teacher response to classroom reform suggests that teachers can respond to policy interventions in a variety of ways. However, the conditions of the classroom and the work of teachers creates an atmosphere in which those tasked with enacting simultaneous and sometimes competing policies from multiple governance levels have little opportunity to understand or realize the original policy intent (Kennedy, 2005). Current evaluation reform, unlike curriculum reform, extends external accountability pressure to the level of the individual teacher. Therefore, it is possible that teachers may react in a variety of ways to meet the requirements of the evaluation policy that may be similar to those demonstrated by teachers under curriculum reforms.

One predicament in teaching is dependence on students to participate in changes in practice (Cohen, 2011). Because teachers are dependent on students' success under current evaluation policies, there are powerful incentives for dramatic changes that can lead to new behaviors, skills, habits, and understandings. Many of these possible behaviors were discussed in the previous section on teacher responses to external accountability pressures. Alternatively, it is also possible that the lack of cohesiveness among schools may lead to teachers perceiving major changes occurring when outsiders actually view the change as minimal (Cohen, 1990). So, literature on teacher response to curriculum reform can help predict ways in which differences in school sites may interact with the evaluation policy to yield different effects across and within sites.

Several frameworks have emerged that identify typologies which teachers exhibit when faced with reforms. One framework that has been utilized when looking at teacher response to classroom reform was employed by Oliver (1991) in describing the strategic processes that organizations employ in response to external pressures. Oliver describes a typology of strategic responses including: acquiescence, compromise, avoidance, defiance, and manipulation. Coburn (2004) argues in her study of the implementation of a reading policy that "the relationship between institutional pressures and classrooms was much more interactive and nonlinear than that portrayed by Oliver. The teachers were connected to messages from the environment via a web of interactive linkages through which messages about reading moved in, out, and around schools through multiple routes" (p. 223). Coburn felt that there were conflicts between her observations and Oliver's views of both denial and acquiescence. As a result, Coburn offered five alternative typologies: rejection, decoupling/symbolic response, parallel structures, assimilation, and accommodation (2004).

Alternatively, a recent piece on educator actions in a competitive marketplace has condensed these typologies into three typologies which were dependent on the perceived legitimacy of the reform: acquiescence, denial, or adaptation (Yurkofsky, 2016). Under this condensed version, those who acquiesce accept the policy and modify practice around it, those who deny it may disregard or revolt against the policies ideals, and those who adapt may try to weld existing practices and beliefs with policy priorities in order to ensure survival in the system.

Regardless of the specific typologies used, the general idea that teachers may perceive evaluation policy legitimacy in varying ways and act according to their perceptions is relevant to this proposed study. These perceptions may differ based on the perceived legitimacy of the policy within the school site, the teacher's relative security in their job, and the usefulness of the feedback received, points which all emerged in interviews with focal teachers. So, with this in mind, I have adopted Yurkofsky's three typologies, which were designed to focus on educator actions in competitive markets, to my dissertation.

Leadership Capacity and Evaluation

There is emerging evidence that school leadership impacts the success of evaluation policy both in terms of implementation and in the quality of feedback in which teachers receive. In the case of North Carolina's policy, school administrators are the individuals who conduct most formal evaluations. Researchers have documented that the roles of principals have shifted over time to include an expanded role as an instructional leader due to changes in both policies and public expectations (Bryk, Sebring, Allensworth, Luppescu, & Easton, 2010; Louis, Dretzke, & Wahlstrom, 2010; Spillane & Kennedy, 2012). Additionally, policies are subject to interpretation and alteration by those who are tasked with enactment in real contexts, resulting in what Lipsky termed "street-level bureaucrats" (2010). Because of this, the capacity of administrative leaders to conduct evaluation can impact the way in which evaluation policy is implemented in varying school contexts as evaluating principals become street-level bureaucrats of the policy.

For instance, studies have uncovered some of the unintended consequences of having principals conduct formalized evaluation. First, principals have varying views on both the purpose and the use of evaluations and may respond to one policy message at the expense of others, leading to varied implementations of the policy (Kraft & Gilmore, 2016. Reinhorn, et. al, 2017). Additionally, the aforementioned expanded role of principals has contributed to a deficit of time to devote to evaluation. Furthermore, a lack of experience in the subject area being observed may result in narrowed feedback being provided to a teacher that does not allow for improvement of instructional practices (Kraft & Gilmore, 2016). Studies have also suggested that the quality of feedback a teacher receives from an evaluation is dependent on principals having the necessary training, time, and resources to devote to provide individualized, actionable feedback (Kraft & Gilmore, 2016. Reinhorn, Johnson, & Simon, 2017). Similarly, principals who are well versed in the application of good instructional practices are best prepared to engage teachers in a process of inquiry, reflection, and improvement (Reinhorn et al., 2017). The unintended consequences of the school-level administrator's role in implementing evaluation policy contributes to variability in the success of the policy across school sites.

Studies have also demonstrated that principals may assess teachers differently on formal evaluations when opposed to summative evaluations. Two recent studies demonstrate that while principals still tend to overall evaluate their teachers quite positively, more positive ratings tend to be assigned on high-stakes assessments versus low-stakes assessments, and principals verbally report ineffective teachers in their school despite formal evaluation ratings demonstrating otherwise (Grissom & Loeb, 2017; Kraft & Gilmour, 2016). Furthermore, these differences are also amplified when the stakes are higher for individuals. For instance, new teachers, who have limited career protections and are therefore more likely to be adversely affected by a negative evaluation than an experienced teacher, are often rated much more positively on high-stakes assessments versus low-stakes (Grissom & Loeb, 2017). This demonstrates that principals are reluctant to show criticism on formal evaluations that may be expressed in lower stakes situations.

There are a few possible explanations for the variability in ratings given on high-stakes assessments versus low-stakes assessments as well as for variability across experience groups. First, it may be possible that principals find more value in providing formative feedback to their teachers through informal means versus using high-stakes, summative evaluations. For instance, in a study of six schools, all of the principals interviewed began referencing their approaches to formative evaluation rather than summative evaluation, suggesting that formative, low-stakes feedback may be more valued by administrators (Reinhorn et al., 2017). Additionally, principals in one study cited time constraints as a reason to be more lenient in high-stakes evaluations as they were unable to provide concrete feedback to improve practice (Kraft & Gilmour, 2017). Principals also explained that they wanted to recognize teacher effort and to evaluate potential while simultaneously motivating teachers towards achieving that potential (Kraft & Gilmour, 2017).

Additionally, principals may rate their teachers in a particular way in order to protect their staff. In one study, school administrators expressed concern over the difficulties of replacing a teacher who was either removed or felt pressured to remove themselves from the classroom due to poor ratings, particularly with newer teachers who may not have any career protections (Grissom & Loeb, 2017; Kraft & Gilmour, 2017). Differences in how experienced versus non-experienced teachers are evaluated were also found in one study of all states plus DC and 25 large school districts (Steinberg & Donaldson, 2016). Overall, the reluctance of administrators to critically evaluate teachers on high-stakes assessments suggests that principals may be attempting to protect staff from the consequences of low scores or otherwise feel unable to be as critical as they would be in low-stakes situations.

Theory Driving Research

The preceding literature review informs the building of a theory for investigations into how teacher evaluation may impact classroom practices (Figure 1). The theory is that motivation and feedback provided by evaluation are factors that interact and create an impetus for action on the part of the teacher. However, both motivation and feedback are filtered through various aspects of teaching conditions. For this study, the school-level factors that will be examined as part of teaching conditions include evaluation conditions and evaluation status, and individual-level factors include years of experience, licensure, and subject area. According to the proposed framework in Figure 1, these factors filter the policy to yield classroom practices. There are other potential factors that can affect classroom practice, but these two school-level and three individual-level factors remain the focus of my study.

According to this theory, teacher motivation associated with evaluation is influenced by both extrinsic and intrinsic rewards. Additionally, conditions are informed by school-level factors (evaluation conditions at a school and the existing evaluation status of teachers at a school) as well as individual-level teaching conditions (experience, licensure, and subject area). Finally, while we do not yet know how teachers may specifically react to evaluation in regard to classroom practice, existing literature on teacher responses to accountability pressures and to classroom reform predict the ways such reactions may manifest. This study was designed to specifically examine the extent to which teachers felt their practice was influenced by evaluation with particular attention to modifications in what is taught, the teaching strategies utilized, and the directing of focus on certain students based on evaluation. I was unable to gauge whether turnover was related to evaluation, but I did ask questions that gauged teacher perceptions of evaluation as related to their perceptions of fairness and job security.

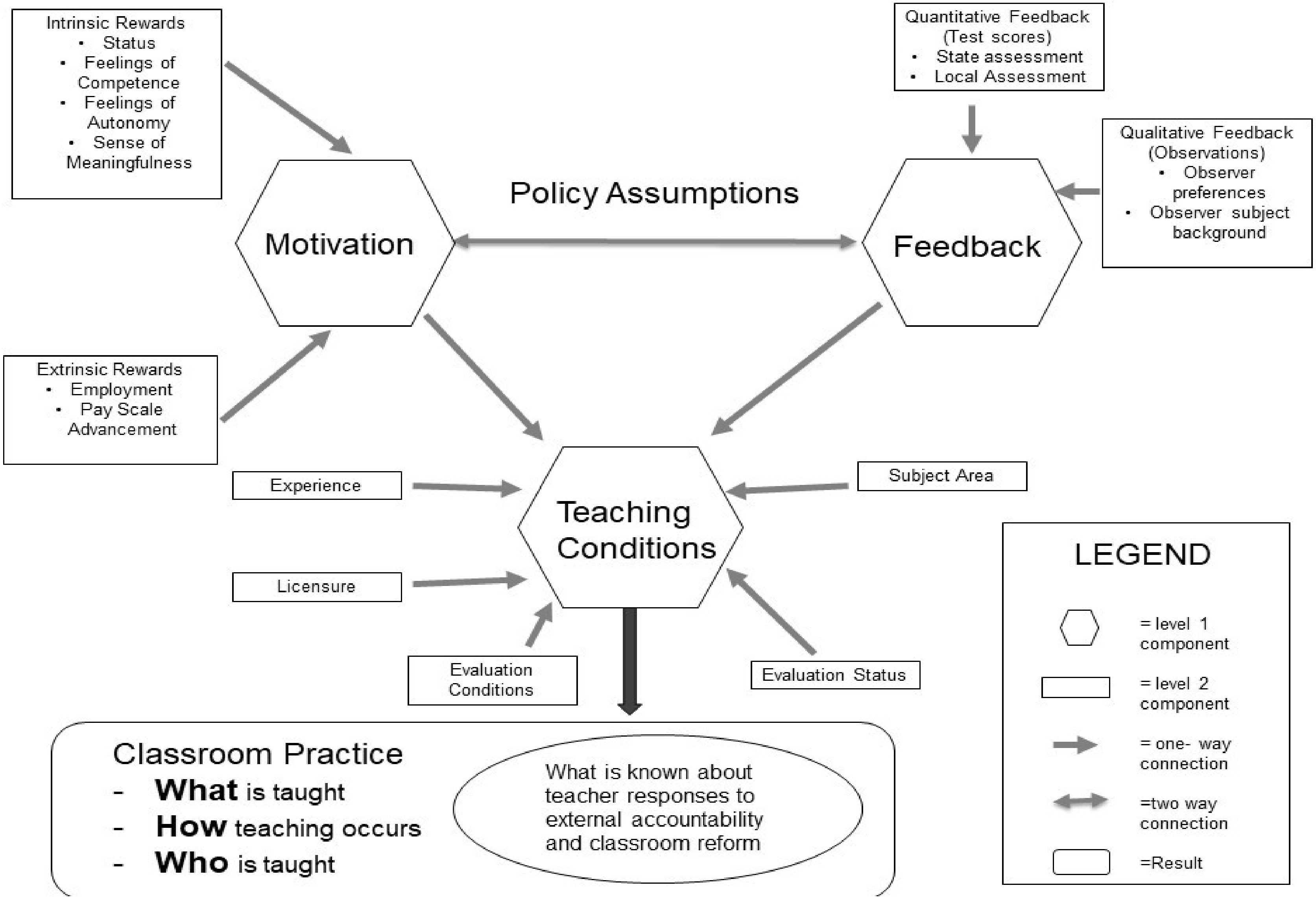

Figure 1. Framework to Guide Research on Evaluation and Practice

CHAPTER 4: Research Design and Methodology

The format of this dissertation is a mixed methods case study describing and explaining the relationship between teacher evaluation policy and teacher practice in light of various contexts and conditions. Case studies are an ideal design for attempting to understand a particular phenomenon where multiple variables interact in a single context (Derrington, 2013; Halverson & Clifford, 2006; Miles, Huberman, & Saldana, 2014; Yin, 2009). The scale of this case study is four high schools of varying contexts in a single school system. This scale not only allows for a breadth of analysis across locations in the district, but also for a detailed, in-depth look at how individual teachers see evaluation interacting with their classroom practice. So, I was able to collect data across system, school, and individual contexts. Three major types of case studies are commonly used to study research questions including: exploratory case studies, descriptive case studies, and explanatory case studies (Berg, 2007). My research questions focus on describing relationships within a phenomenon and, when possible, explaining what influences individual behavior in a case. Therefore, this study meets the criteria of both a descriptive and an explanatory case study.

Furthermore, mixed methods are utilized to answer the research questions in this dissertation. Mixed method research can be formally defined as "the class of research where the researcher mixes or combines quantitative and qualitative research techniques, methods, approaches, concepts, or language into a single study" (Johnson & Onwuegbuzie, 2004, p. 17). Statewide, publicly available quantitative data were used in the selection of the research sites. Additionally, the data was collected in three phases. The first phase of survey collection represents the quantitative stage, though open-ended commentary was also permitted to allow survey participants to explain answers more fully if they desired. The second and third phases consisted of interviews which were analyzed qualitatively to better explain the findings from the survey. In this manner, the qualitative data also served as a check for the quantitative analysis. Furthermore, the approach taken to analysis allows the quantitative data to describe what is happening while the qualitative work helps explain the phenomena.

Johnson and Onwuegbuzie (2004) contended that the objective of mixed methods research is to draw from the strengths and minimize the weaknesses of qualitative and quantitative methodology which results in research which is superior to that conducted with one method. Because case studies allow for a nuanced understanding of the particularities of context and mixed methods studies allow for analysis which can address my research questions more fully than a single method approach, the surveys and interviews, along with publicly available district- and school-level data, allow me to effectively address my research questions by both describing and explaining the relationship between teacher evaluation and practice at the four study schools.

Participants and Sampling Strategy

Participants in this study are high school teachers (N= 45) in North Carolina. The focus on high school teachers is important for two reasons. First, the subject area distinction is more pronounced at this level (compared to elementary-level teachers) because high school teachers usually hold degrees in the subjects they teach rather than broadly in education (that sometimes include a major, but not a degree in a subject area). The teaching certificate is often secondary to the subject degree in North Carolina. Secondly, most North Carolina high schools, including the four in this study, follow block schedules where courses are taught over half a year and then change for the second semester. The block scheduling allowed me to conduct follow-up interviews after a semester-long course had ended, students had taken assessments, and teachers had an idea of how their evaluation was going with respect to the student growth score they may receive.

The selection of high schools for my study involved examining district level data. First, I reviewed the North Carolina Working Conditions Survey results and Educator Effectiveness results for each high school to determine evaluation conditions and effectiveness status. These sources are described in greater detail in the next two sections. I then identified four focal schools for the study that fit varying combinations of high/low evaluation conditions and high/low effectiveness status. I describe these measures in the next two sections of this chapter.

Teacher Working Conditions Survey and Evaluation Conditions

The Department of Public Schools of North Carolina, in conjunction with the North Carolina Association of Educators, administers a Teacher Working Conditions Survey (TWC) biannually which asks teachers to answer questions about varying aspects of their working conditions, including topics such as professional development, facilities, and community support. The overall response rate in Broadville County for the 2015-2016 school year was 79.79%. The data from these surveys are publicly available (http://www.ncteachingconditions.org/). There are nine questions on the survey which focus specifically on evaluation (Table 2). Seven of the questions are directed towards local assessment, such as observation, either by explicitly stating the focus is local or by being components of a larger section on local conditions. Two of the questions focus on testing, which is the state level component of evaluation. These questions were used to determine the evaluation conditions of individual schools in a method I will next describe.

Table 2

Teacher Working Conditions Survey Evaluation Related Questions

7.1d Teachers are held to high professional standards for delivering instruction.

7.1f Teacher performance is assessed objectively.

7.1g Teachers receive feedback that can help them improve teaching.

7.1h The procedures for teacher evaluation are consistent.

9.1a State assessment data are available in time to impact instructional practices.

9.1b Local Assessment data are available in time to impact instructional practices.

9.1c State assessment accurately gauges students' understanding of standards.

I used data from the latest administration of this survey (Spring 2016) to determine a school's evaluation conditions (North Carolina Teacher Working Conditions, 2017). The original survey responses are presented in a Likert-type format; however, the data is also reported as a percentage of the total number of people who indicated any level of agreement. As previously mentioned, some of the evaluation-based questions focused on the local level and others at the state level. There were some drastic differences among the scores for locally focused questions and state focused questions, so I separated the questions based on whether there was a specific state or local focus to create two distinct scores, one for local and one for state. I then created composite averages of the percentage of respondents who indicated some level of agreement separately for the state and local categories. For each high school, I compared the scores of each of the aforementioned categories and measured the distance of the school's percentage from the system's average. This calculation yielded either a positive or negative number which indicated distance from the system mean. These numbers provided a school's condition score for each category (Table 3).

Table 3

Calculating Condition Score

	Teacher Working Condition Survey Questions											
Location	7.1d	7.1f	7.1g	7.1h	9.1a	9.1b	9.1c	Local Score Composite	Local Condition Score	State Score Composite	State Condition Score	Response Rate
Broadville	92	84	81	83	52	75	31	83		41.5		79.8%
Riley	94	83	90	75	48	70	33	82.4	0.6	40.5	-1.0	94.1%
Phoenix	100	94	88	95	29	60	22	87.4	4.4	25.5	-16.0	63.3%
Charles	88	95	78	81	39	63	22	81	-2.0	30.5	-11.0	90.0%
Central	90	72	72	84	22	49	16	73.4	-9.6	19	-22.5	50.0%

Note. The full text of the survey questions are located in Appendix A

Educator Effectiveness Database and the Effectiveness Score

I used data for the school year (2015-2016) that preceded the study year (2016-2017) from the Educator Effectiveness section of the North Carolina School Report Card database to calculate Evaluation Effectiveness scores. I was also able to separate this score by local and state focus, as I will describe later.

The website for the Educator Effectiveness database states in highlighted text that, "North Carolina's Educator Evaluation System is a ***growth instrument***. It identifies the knowledge, skills, and dispositions expected of teachers, and measures the level at which teachers meet the standard as they make changes to their teaching" (emphasis is consistent with the referenced text) (Educator Effectiveness Database, 2015). The instrument consists of six standards (See Table 4). The website also specifies that due to teachers and administrators being lifelong learners, "It is ***expected that teachers in a school would be distributed across the rating categories***" (emphasis is consistent with the referenced text) (Educator Effectiveness Database, 2015). The first five of the six standards of the evaluation instrument debuted during the 2010-2011 school year. Legislation current at the time of writing states that career status teachers must receive a full evaluation of all six standards at least once during a five-year license renewal cycle. Otherwise, career status teachers can be evaluated on an abbreviated cycle. All teachers who had not received career status prior to the 2013-2014 school year are subjected to the full evaluation cycle each year.

Standards 1-5 are observation standards determined locally by school-level administration with five possible proficiency ratings, whereas standard six is based on student growth data on state exams, determined by a state software system, and has three proficiency levels (Table 4). In the past, the proficiency for standard 6 was determined at the high school level by individual student test data for teachers of the three state tested subjects: Algebra II, Biology, and English II. At the time of this dissertation, those results are used for schoolwide scores which are combined with individual teacher scores from students taking the North Carolina Final Exam in order to calculate a teacher's standard 6 score. At the time of this study, the North Carolina Department of Public Instruction had announced that standard six was going to be "devalued" and spread across the other five standards; however, at the time of writing it was unclear how that would occur. The methodology and technology for calculating standard six scores as well as the assessments used to determine such scores are all conducted by the state. Again, in the 2013-2014 school year, North Carolina removed career status as a designation obtainable by teachers who had not yet received it. It is notable that teachers lose career status if they switch between systems in the state and may have been unable to retain that status. Non-career status teachers are on a one-year contract structure and must be evaluated on a full cycle every year indefinitely regardless of the years of experience.

Table 4

Evaluation Rubric

Evaluation standards for teachers		Type and Ratings
1	Teachers demonstrate leadership	**Observation (Local)** Not Demonstrated Developing Proficient Accomplished Distinguished
2	Teachers establish a respectful environment for a diverse population of students.	
3	Teachers know the content they teach.	
4	Teachers facilitate learning for their students.	
5	Teachers reflect on their practice.	
6	Teachers contribute to the academic success of their students	**Student Growth (State)** Does Not Meet Meets Exceeds

Evaluation Effectiveness scores were calculated using data from the 2015-2016 school year, which is the year that preceded the study year. Standards 1-5 were not applied to all teachers as those with career status could be evaluated on an abbreviated evaluation schedule at the discretion of the observing administrator. So, I could only calculate the average number of standards proficient, not an average of teachers who were proficient for standards 1-5. For standard 6, the number of standards and number of teachers are the same. First, I calculated an average of standards proficient for standards 1-5 by summing the total number of proficient counts for all five standards and dividing that by the total count for standards 1-5. Standards 1-5 are awarded locally by school-level administration following observation and are labeled as "local" scores. Standard six was more straightforward as it was calculated by the state based on standardized student assessments. I summed the number of teachers who met the standard and divided that by the total number of teachers. I then took the averaged percentages for each school and subtracted each school's average from the system's average to create Effectiveness Scores for each school for both local and state measures (Table 5).

Table 5

Establishing an Evaluation Effectiveness Score

Location	Local Proficient	Local Score	State Proficient	State Score
Broadville	99.0%		88.0%	
Riley	99.6%	+0.6	89.5%	+1.5
Phoenix	100%	+1.0	75.0%	-13
Charles	96.5%	-2.5	97.3%	+9.3
Central	100%	+1	92.2%	-4.2

Phase 1: Surveys of Sample Schools

For the first phase of research, I administered a survey to Mathematics and English teachers at the focal schools to identify ways in which evaluation influenced teacher practice during the previous school year as well as the anticipated effect on the upcoming year. The first series of questions on the survey were demographic questions designed to identify years of experience, licensure type, what subjects a teacher had taught, past and current status as a teacher of tested or non-tested courses, and current status as a teacher of End of Course (EOC) or North Carolina Final Exam (NCFE) tested courses. In the demographics section I replicated the nine evaluation condition questions from the TWC survey to establish a measure of the individual teacher's satisfaction with the conditions at the school. This helped me identify whether or not a teacher deviated significantly from school-wide responses and assisted in my selection of focal teachers.

The survey then included Likert-scale questions requesting that teachers reflect on their prior year including: the extent to and way in which evaluation affected their motivations to succeed in the classroom, their use of feedback from evaluations, as well as their perceptions of job security and accuracy of the evaluation, and the ways in which evaluation guided what was taught, how it was taught, or on whom focus was directed in the classroom.

The third portion of the survey asked the same questions about anticipated behaviors "looking ahead" in the new school year and how teachers planned on modifying practice in the current school year. The final two question sets were complementary and are referred to as the "complementary question set" throughout the dissertation.

During an initial analysis of this survey, I identified focal teachers at each school and attempted to procure two teachers from Math and two from English to participate in the interview phase**.** Descriptive statistics and paired t-tests were run on the survey responses first as a whole sample, then by school-level, and then by individual-level characteristics.

Survey instrument. The survey consisted of three sections and is available in Appendix A. The first section asked participants for demographic data. The second section replicated nine questions from the Teacher Working Condition Survey that were used to calculate the school Evaluation Condition scores as described prior. The final section contained a complementary question set that asked teachers to reflect on the previous year and then the current year.

The nine questions from the Teacher Working Conditions survey used to determine Evaluation Condition Scores were replicated on the survey administered in Qualtrics to get a sense of the perceptions that Math and English teachers from the focal schools had of evaluation conditions. I used the same scale that was used in the original state-administered Teacher Working Conditions Survey which included the options “Strongly Disagree,” “Disagree,” “Agree,” “Strongly Agree,” and “Don’t Know.” To analyze the results of this section for Chapter 5, I eliminated the “Don’t Know” responses question by question which resulted in a different reported N across questions.

The bulk of the survey featured questions asking teachers to reflect on the previous school year and then a complementary set of questions asking them to think about and anticipate the current school year. Each question set had a unifying theme. However, one question about the prior year was not replicated in the current year question set; that question asked teachers to evaluate the statement: “Last year’s evaluation will impact decisions about classroom practice in the upcoming school year.” The nature of this question did not allow for a complementary question in the second set. Also, the number of participants in each set of the survey is different because some participants were not in the classroom in the prior year and therefore the section reflecting on the previous year was not applicable to those individuals. Additionally, responses were not forced, so some participants opted not to answer all of the questions, which also contributed to variations in the N question by question. Teachers were asked to evaluate all of the statements using the following Likert-type scale, where the higher numbers indicate a higher level of agreement: 1. Strongly Disagree; 2. Disagree; 3. Neither Agree nor Disagree; 4. Agree; 5. Strongly Agree.

Phase 2: Preliminary Interviews of Focal Teachers

The first round of interviews was conducted two months into the 2016-17 school year. The purpose of these interviews was to better distinguish the relationship between teacher practice in the classroom and the evaluation policy. I attempted to sample two English and two Math teachers from each school. However, I was unable to achieve uniform sampling across subject areas. Riley did not have any Math teachers who were willing to be interviewed and Phoenix only had one Math teacher who was willing to be interviewed. Conversely, at Central the English Department Chair recruited teachers for interviews, and due to a communication error, selected three English teachers to be interviewed.

During the first interview, I tried to identify typologies of reform response from teachers as well as to parse out differences between individuals of varying characteristics. First, I asked teachers to generally explain their experiences with evaluation both in the past as well as so far in the current school year. The next questions asked during the interviews were developed based on both the school-based and individual-level responses to the survey items. This was done in an attempt to find explanations for differences that were related to the context and conditions of specific teachers. Finally, I ended each interview by asking every teacher their thoughts on the two policy assumptions of evaluation: (1) evaluations are necessary because teachers need to be rated, sanctioned, or rewarded in order to be motivated to do a better job; and (2) evaluations yield information that is useful for teachers to improve practice.

Phase 3: Follow-up Interviews of Focal Teachers

Follow-up interviews of the focal teachers were conducted in mid-March of the 2016-2017 school year. At that point, teachers had been through one state testing cycle in January for the first semester. Due to the block schedule system used in all the study schools, teachers were teaching entirely new courses. At this point in the year, every teacher had been evaluated at least once and nearly all of them had completed all the required evaluations for the year. I began the interview by asking teachers for an update on their observations for the year. I also asked teachers how testing had gone and if they had any surprises from the process or the scores. I inquired about teachers' courses in the current semester and if they felt any differing pressure from state testing with the courses they had currently versus the prior semester. I focused on attempting to identify any changes in typology, perception, or behaviors based on the evaluation in the first half of the year. Finally, I shared with each teacher the status of their school's Evaluation Conditions and Evaluation Effectiveness scores as I had calculated previously. I asked teachers if each specific score surprised them or if they thought it was an accurate reflection of the climate of their school and why. I also asked teachers to reflect on if anything had changed in the current school year that may alter the scores if this study were to be replicated with similar data from the current year. Aside from serving as a new source of data, this interview also served as a member check to ensure validity of the study (Deyhle, Hess, LeCompte, 1992).

Data Analysis and Establishing Validity

Quantitative

The survey data served three purposes in this study: as a source of data for analysis, as a mechanism to identify focal teachers for the interview portion, and to provide information used to develop individual-level questions for the interview phase. All quantitative analysis of the survey data was completed using SPSS software. The data were analyzed in three ways: as a whole sample, at the school-level, and based on individual teacher characteristics. I first conducted a sample-wide analysis of the data. This analysis included calculating descriptives and conducting paired sample t-tests to identify differences between the prior year and the current year for the whole sample.

To examine school-level differences, I conducted two types of analysis looking for differences between schools as well as within schools. First, I calculated descriptives and conducted ANOVA to identify differences between schools for both the prior and current year question sets. I then conducted paired t-tests within each school to determine differences within each school for responses on the prior year versus the current year.

I examined three individual teacher characteristics in the survey data: licensure, seven-year status, and subject area. First, descriptives were calculated for each of the three categories for both the prior and current year question sets. Then, independent sample t-tests were conducted to determine differences between the categories of each characteristic on both the prior and current year question sets.

Qualitative

Both the preliminary and follow-up rounds of interviews were audio recorded, transcribed, and checked for accuracy. Copies of the transcripts were provided to the interview participants so they could ensure that the interview appropriately reflected their intended meaning. I organized and coded all of the interview transcripts using the qualitative data software, Dedoose.

A coding scheme was developed inductively. The coding scheme would be considered open-coding as the codes developed as my work progressed rather than being pre-determined outright. However, most of my codes were grounded in the results of my literature review. Specifically, I focused on the different types of motivation (extrinsic and intrinsic), the types of responses teachers demonstrate in research on accountability pressures and classroom reform (acquiescence, denial, and adaptation), and some of the types of reform responses teachers engaged in (selecting curriculum, selecting teaching strategies, and directing focus on students). I started with these aforementioned grounded codes and developed new codes and child codes as trends further emerged. In this manner, coded material was grouped together by emerging theme and typologies. Codes were not mutually exclusive. The validity of my codes was confirmed by double coding 36% of the data, and any discrepancy was noted and addressed in order to look for alternative interpretations of the data (Miles et al., 2014). Descriptions and examples of interview codes are available in Table 6.

Table 6

Interview Code Descriptions

Codes	Definition	Example
Motivation	Reference to being motivated to better perform in the classroom, better teaching practice, increase student achievement, or increase performance in some other aspect of the teacher's job.	"Teachers do not need to be ranked in order to be motivated to do better. I feel like one doesn't enter teaching for that. The people who are entering teaching are doing it for intrinsic motivations because they generally want to help, and that competitiveness just takes away from the whole goal, which most teachers have which is to help students learn… I guess what motivates me is students having curiosity, and the pursuit of intellect, that motivates me."
Internal	Reference to a form of internal motivation. This reference may include a teacher being motivated by disappointment or achievement in the evaluation process or experience.	"I would say that that probably comes down to why someone came into the profession in the first place. I do not feel like I need affirmation from my Principal as much as I do feel like I really actually care about my students' growth and learning. And so, I came to be an educator simply because I believed in the ability to make an impact in this world, and I see the need for it… And I think that probably drives me forward more than anything else. I think that also, you have to really love what you are teaching and the process, right? Because I do think it is a hard profession, and it kind of beats people down really quickly. And without that motivation or that affirmation, I think a lot of people do get lulled to sleep a lot."
External	Reference to a form of external motivation including things such as pay increase and the achievement of ratings on an evaluation rubric.	"If I had an evaluation score that was really low, that might motivate me to see, 'What did I do wrong?' And, 'Let me try to do better.'"

Table 6 (cont'd)

Observation Feedback	Reference to feedback gained from the observation process or rubric.	"I try and find value in them and I think a lot of times I will get out of them something different than what I expected to, like 'Oh hey, I noticed this and this in your classroom and I'm wondering if you might try this idea or this type of formative assessment' or something that has been really helpful but wasn't necessarily what I expected going in and so I'm wondering if, with teachers that are a little more seasoned, that if they have those little things, because I feel like a lot of times that advice that I'm getting is maybe like "Oh, I might have figured this out in a year or two." And so, I'm wondering if you've got all of that, because there's only so much you can see in 45 minutes or an hour, hour and a half. That if there is anything you can see, it's probably one of those things that's pretty easy to fix."
Negative	Reference to feedback which describes the evaluation process, feedback, or scoring in a negative or detrimental manner. May also highlight aspects of the feedback which teachers perceive make it not useful.	"I remember a couple years back, they started putting all that data into EVAAS for us to look at, from school to school. You could look at the different schools and just see the different standards …and you could just see what the average evaluation score was in each category, in each school. And there were some schools that were just consistently, much higher. It was odd, and I cannot remember which school it was, but you look at one school and 45% of their teachers had the highest score in almost every single category. And you look at another school, it is 20 minutes down the road, and 10% of their teacher had the highest score in every category. The problem, and I do not know if this is a training thing for the administrators, or if it's a systemic thing, what it is…And we all looked at that and said, 'Well, this doesn't make any sense, if Principal A over at that high school's just going to get everyone a five just because either they have low standards or they're evaluating based on just the talent they have and maybe they do not have a real strong talent pool. That removes a lot of the objectivity... Because the same person's not doing all the observations and they are not

Table 6 (cont’d)

		holding everyone to the same standard, there is going to be a problem. We even talked about that in the schools, we know that sometimes if you get a certain administrator assigned to you but that means, ‘Oh, yeah, my scores are going to be great.’ Because for whatever reason that person, they are busy doing other stuff, they have multiple things they're dealing with or they just are more laid back or easy going, or sometimes they just do not have the same time in the classroom to know what they are looking for all the time. And the standards aren't the same all the time.”
Positive	Reference to feedback which describes the process, feedback, or scoring in a positive or helpful manner. May also highlight aspects of the feedback which teachers perceive make it useful.	“The feedback that I get or that I have gotten in the past from evaluations has often been... very specific because when you only see a small snippet of someone's classroom or their teaching style or whatever that day happened... I think it's most effective if you focus on, ‘I solve this one thing specifically. And if it comes up again or when it comes up again this can help.’ And so, the last observation I had last year, the Principal was in here and she was watching, and she gave me a suggestion where I would asked students... I gave them the, ‘Does everyone understand?’ We did my five seconds of nodding and looked around and try to make eye contact with everyone. And afterwards [the principal] said, ‘That was good. Try this’ and gave me a list of three or four different little quick snap formative assessment like, ‘Everyone put your head down. Give me one, two or three.’ And now that's what I do, and I feel like it informs my teaching much better than what I was doing which was just a very simple glance around try and read everyone's face. Those specific things more so than any big grand teaching strength or weakness that I might have that is really hard to observe in 40 minutes or 50 minutes.”

Table 6 (cont'd)

Testing Feedback	Reference to feedback gained from the testing process, scoring, or score reporting.	"Well, so when we get our feedback on test results from the state, it's divided basically into three categories. RL, so Reading Literature, RI, Reading Informational text, and then Language, which is vocabulary skills. And that is the only breakdown of that test data that we get, are those three categories. So, within reading fiction there are so many components of reading fiction, so I never have any idea if my students are struggling with characterization or plot structure. Or I never have any idea if with RI, if they are struggling with central idea or supporting details. So, there is no way for me to use that data to actually improve my instruction other than if I was weak in informational. Let us try to throw more non-fiction in… so there's no specific feedback for me to build on."
Negative	Reference to feedback which describes the evaluation process, feedback, or scoring in a negative or detrimental manner. May also highlight aspects of the feedback which teachers perceive make it not useful.	"So, I feel like we are doing a great job here. But the test, if they are just looking at achievement, I do not think that shows everything. We are looking at growth. We're doing pretty darn well. But in terms of the test's usefulness, its effectiveness at determining what our kids know, I don't feel that it does overall. I just do not think you can accurately gauge what students have learned in a 90 day course or 180 days, on a 30, 40 question test, especially one that is multiple choice, at least, part of it is multiple choice.
Positive	Reference to feedback which describes the process, feedback, or scoring in a positive or helpful manner. May also highlight aspects of the feedback which teachers perceive make it useful.	"The kids that I actually saw growth from in my class were those kids that I saw growth from on the test."
Work Decisions	Reference to changing some aspect of teaching in anticipation of or the result of an evaluation.	"Our big thing in the past couple of years has been learning targets. They want everyone to have some kind of learning target on the board. And that is not something that I have really done before. I

Table 6 (cont'd)

		always have an agenda on the board of, 'Here is what we are doing today,' and as we're going along we talk about, 'Why we are doing those things.' But I never specifically said, 'I will be able to do this,' or 'I will be... 'I never had that goal, stated in that way. So yeah, that's something that I'll put up there now because that's just something that on the observations they told us, 'We are going to look for these learning targets.' And so, I will make sure that I am putting them up there, even if I don't always agree with the whole process because I know that's something they are looking for."
Strategy/How Taught	Reference to choosing or altering a teaching strategy in anticipation of or the result of an evaluation.	"Some of it helps to me to change minor things in my instruction, a little bit. But it is usually... What I mean by that is if students see an equation written a certain way, I know to make sure to show them that format versus another format that's not incorrect... I would not be teaching an incorrect format."
Curriculum/What Taught	Reference to choosing, excluding, or otherwise altering curriculum in anticipation of or the result of an evaluation, which may include the teaching explicit testing strategies.	"The pressure is that the Math One does have the end of course test. [Three teachers] developed a plan together to create a spiral review throughout Math One. And it has been going so well because each day of the week we have a different type of warm up activity, and we're reviewing a specific outcome that they covered last semester in foundations to Math One. And now we are also beginning to review the ones that we started at the beginning of this semester. But it's really great for us to pick up on little details, that we are like, 'Oh, is that what they were missing? They could not tell the difference between a solid line and a dotted-line graph?' Who knew that was the little missing piece of information? And it has not been perfected yet, but it has been really helpful for us."
Who is Taught	Reference to directing or not directing focus on certain students (triaging) in anticipation of or the	"What happens, I think, in my mind is, which can be a dangerous thing, sometimes I think, when I get the Honors class, I naturally expect that they will do fine on the exam. And so, where I spend

Table 6 (cont’d)

	result of an evaluation, which could occur within classes or across classes.	a lot of time with the Standard class test prepping, I do not spend as much with the Honors class. And they usually do fine. Their growth is usually not as large, which it is harder to meet growth anyways. Their test growth is usually not as good. Even though they meet what they should meet to pass, they don't grow as much. But my bigger concerns for the Honors class shift more toward writing, which we are not evaluated on at all, that writing prep that they need for the college-level writing, and just those critical-level thinking skills, the research skills, some of those bigger things that I can spend more time with them in the Honors class and really do not get a chance to go into with the Standard class because we are test-prepping.”
Response to Reform	Teachers make a statement that exhibits an adherence to one of three reform typologies identified in this dissertation.	“Nobody cares if you actually teach what you are supposed to teach. They are just glad you showed up… I know what to do. I have got the degrees. I know what to do. You do not need to be constantly telling me what to do. And because they do not invade our space very often, who knows?”
Acquiescence	A statement that reflects a typology where the individual accepts the policy without question and feels the policy did not impact their lives or jobs.	“I think teachers' attitudes toward the observation of fairness, none of us feel like we are targeted or there is pressure put on us or anything like that. At the same time none of us, I do not think, feel like we are getting amazing feedback for growth and whatever… it is not impacting us one way or the other, we are just doing what we do in our classrooms every day.”
Adaptation	A statement that reflects a typology where the individual adapts the policy to fit their own needs.	A teacher describing how she uses evaluation for self-assessment: “I feel like, when I go through the standards in that pre-evaluation is when I learn the most about, ‘Am I doing these things? Which of these could I do more?’ I can talk about curriculum and classroom management with my instructional coach, but these other things, like, ‘Am I contacting parents?’ For me, when I read through and

Table 6 (cont’d)

		did my pre-evaluation this semester, I was like, ‘Oh, I am so bad at that. Maybe I should work on that a little bit.’ So, when I looked at that pre-evaluation, and I kind of looked at those things, they were asking me like, ‘Do you contact parents regularly and stuff?’ That is our standard or something like that. I was like, ‘I could do better at that.’”
Denial	A statement that reflects a typology where the individual openly rejects or rebels against the policy due to a perceived negative impact.	“The principal tells people, ‘These are what your PDP goals are going to be.’ I was like, ‘Are you kidding me? You are not telling me.” I say, ‘I am going to do what I want to work on. I am not going to work on what you tell me to, just because you just told me. I am going to be that bad kid.’ Because, I believe that I should be free to pick my own things to work on.”

Researcher Background and Neutrality

Aside from the double coding of data, my study incorporates various internal and external supports for establishing validity. First, I was well-prepared to approach this type of research due to my background as a National Board Certified high school teacher with five years in public schools, including four years of experience in the state of North Carolina. I was teaching in North Carolina when the statewide observation system was adopted and when the student growth standard was added. This allowed me to think about my own experiences during that time to anticipate how the policy may have impacted teachers. My experience also granted me greater awareness of the policy atmosphere in which I was investigating and allowed me to more effectively engage with interview participants.

Additionally, my background equipped me to be able to engage in the work that I conducted in this dissertation. I had extensive coursework in both quantitative and qualitative research methods as well as valuable work experience in my research assistantships dealing with both quantitative and qualitative data. This work experience has also involved developing codes from literature reviews to analyze artifacts such as: interview transcripts, observation rater notes, think aloud transcripts, and student writing.

However, I recognize that bias occurs unintentionally and thus I constantly acknowledged how my past experiences, particularly as a teacher in the same state as my study, may have impacted data collection. I ensured neutrality by writing and reviewing my interview and survey questions beforehand to ensure that I asked non-leading questions and allowed opportunity for clarification from participants. I piloted both the survey and interview questions with other North Carolina teachers in different school systems prior to data collection. Also, the use of multiple types of data in the form of surveys and interviews served as a validity check and I used my initial survey data to triangulate the later interview data (Stake, 2004). My study also features a multiple case design by including various groups of teachers (Math and English teachers, provisionally and professionally licensed, teachers from four different school sites, etc.) which enabled me to test my theory that different groups experience accountability pressure from evaluation in differing ways (Yin, 2009). Additionally, the second set of interviews, which was conducted several months after the first round, served as a type of member check to obtain feedback on the themes and typologies that emerged from the first round of interviews (Deyhle et al., 1992). Finally, my research was guided by a capable committee of faculty from two universities in the areas of teacher education, educational policy, and educational administration and I also sought the feedback of other students in the MSU Educational Policy PhD program throughout the dissertation process (Glesne, 2006).

School System Site

Broadville County is a large school system in North Carolina that surrounds a separate city school system. According to a school system profile available online, Broadville ranks in the top 15 of school systems in size of student population, yet ranks 85th in funding out of 115 school systems in the state. Broadville serves just over 25,000 students and the system website states that over 25% of its students live below the poverty line. The schools in this study demonstrate a high rate of students enrolled in the Free and Reduced Price Lunch program. At the time of this study, there were 23 elementary schools, three intermediate schools, seven middle schools, six regular high schools, one alternative high school, and two middle/early colleges. According to the school system profile, as of the 2012-2013 school year, 14% of students were classified as Exceptional Children (EC), which is North Carolina's designation for those receiving special education services. In the same year, 16% of students had the designation of being Academically or Intellectually Gifted (AIG). Additionally, there were about 15,000 students classified as English Language Learners (ELL) who spoke 66 different home languages.

The school system online profile also states that Broadville employs about 4,000 people and is the second largest employer in the area. According to the NC School Report Card, in 2012-2013 about 20% of teachers had less than four years of experience while an additional 20% had between five and nine. These averages are comparable to the other large school systems in North Carolina.

Broadville was selected for this dissertation primarily due to its size and diversity. A large district was needed in order to be able to identify enough high schools with varying Evaluation Condition and Effectiveness scores to conduct analysis of differences at the school-level. Additionally, teacher evaluation was a sensitive topic at the time of this study and many school systems were facing lawsuits over the implementation of the policy. Increased pushback on state level teacher policies, including the evaluation policy examined in this study, was occurring from the local governments, universities, teacher unions and groups, and the public. Therefore, it was important to be able to provide relative anonymity for the participating district, schools, and teachers. A large district with demographics similar to other, large school districts was necessary to meet such requirements. So, Broadville was an ideal location for this study because of the varying characteristics between its nine high schools and demographics that were similar to other large school systems in North Carolina.

School Sites

Table 7 provides demographic information of the four focal schools in this study. Data on the student population are derived from the National Center for Education Statistics (NCES) Database, which is drawn from the 2013-2014 school year. The teacher data and classroom data come from the publicly available NC School Report Card which uses information from the 2012-2013 school year. The Conditions and Effectiveness Scores were derived from the NC Teacher Working Conditions Survey and the Educator Effectiveness Database, respectably, and were calculated in the manner described earlier in this chapter from data from the 2015-2016 school year.

Table 7

School-Level Demographics

	Riley	Phoenix	Charles	Central
Student population[1]	1591	134	789	1103
% White Students[1]	68%	71%	82%	86%
% Hispanic Students[1]	12%	8%	8%	8%
% Black Students[1]	11%	10%	4%	1%
% Asian Students[1]	3%	1%	1%	1%
% Native American/Pacific Islander Students[1]	0%	2%	1%	0%
% Mixed or Other Races[1]	6%	7%	6%	4%
Students Participating in Free or Reduced Price Lunch[1]	38%	87%	46%	39%
Classroom Teachers[2]	100	18	56	64
Teachers Fully Certified[2]	95%	89%	93%	92%
% Teachers with advanced degrees[2]	36%	56%	31%	30%
% National Board Certified Teachers[2]	24%	38%	41%	38%
% Teachers with more than 10 years experience[2]	63%	53%	67%	66%
Teacher turnover rate[2]	8%	5%	13%	17%
Average English II class size compared to system average[2]	+4	-14	+4	+1
Average Math I class size compared to system average[2]	0	-14	-3	0
Local Condition Score[4]	0.6	4.4	-2	-9.6
State Condition Score[4]	-1	-16	-11	-22.5
Local Effectiveness Score[4]	+0.6	+1	-2.5	+1
State Effectiveness Score[4]	+1.5	-13	+9.3	-4.2
School-Level Growth Score[3]	Meets	N/A	Exceeds	Exceeds

Note. [1] National Center for Education Statistics (NCES) Database, 2013-2014 school year; [2] North Carolina School Report Card, 2012-2013 school year; [3] North Carolina School Report Card, 2012-2013 school year; [4] Calculated as described

School 1: Riley

Riley is the largest high school in the study serving just under 1,600 students and employing about 100 teachers. The student body is the most diverse of the schools in the study with 68% of students being white and 32% non-white. Riley has the lowest level of students participating in free and reduced-price lunch (FARPL) in this study, at 38%. Additionally, 95% of Riley teachers are fully certified, 36% have advanced degrees, 24% are National Board Certified, and 63% have over 10 years of teaching experience. The turnover rate was only 8% and class sizes are reportedly close to the school system's average.

Riley is the only school in this study to have a separate Freshman Academy program geared at ensuring success for students entering high school. Teachers who teach courses for the Freshman Academy are all located on the same wing of the school which is separated from the main body of the school by the cafeteria. Freshman have a dedicated administrator and counselor also located in the wing.

The Condition Score and Effectiveness scores were also quite close to district average. The Local Condition Score was only 0.6 above and the State Condition Score was 1.5 above the district average. Riley was also closer to average on the Effectiveness Scores than any other school at 0.6 above the local and 1.5 above the state. Overall, Riley has conditions and effectiveness that are quite close to Broadville's average.

School 2: Phoenix

Phoenix is the smallest high school in the study, though its population is larger than two other specialty schools in the system. The population fluctuates throughout the year, but it serves about 134 students and employs 18 teachers. It is the second most diverse of the schools in the study with 71% of students being white and 29% non- white. Phoenix has the highest level of students participating in FARPL in this study, at 87%. Phoenix has the lowest percentage of teachers fully certified at 89%, which is possibly an artifact of its small staff size. However, it has the highest percentage of teachers with advanced degrees at 65%, 38% are National Board Certified, and 53% have over ten years of teaching experience. The turnover rate at Phoenix is the lowest in the study at only 5% and class sizes are much, much smaller than the school system's average.

Phoenix is an alternative school that specializes in students who are failing out of or otherwise unable to perform in the traditional high schools. The program is selective and students who want to attend the school must go through an application process to be admitted. The class sizes are quite small which makes Phoenix's alternative education program one of the most expensive programs that Broadville County runs.

Phoenix had the highest Local Condition Score at 4.4 but the State Condition Score was -16 below the district average, indicating a high level of dissatisfaction with state components of evaluation. Similarly, while Phoenix was fairly close to district average for Local Effectiveness Score (+1), the State Effectiveness Score was -13, well below the district average. Overall, teachers at Phoenix have an average to high view of local conditions and an average ranking in effectiveness for local conditions. The State Condition Score and State Effectiveness Score fall far below the system's average. Phoenix serves as an example of a unique working environment with high reported Local Conditions and average Local Effectiveness but very low State Conditions and Effectiveness.

School 3: Charles

Charles is the smallest traditional high school in the study serving just under 800 students and employing 56 teachers. The student body consists of 82% white students and 18% non-white. Charles has the second highest level of students participating in FARPL in this study, at 46%. Additionally, 93% of Charles teachers are fully certified, 31% have advanced degrees, 41% are National Board Certified, and 67% have over 10 years of teaching experience. The turnover rate was only 13% and class sizes are fairly close to the school system's average.

Charles features an initiative to improve Math scores. This initiative involved the creation of a required Introduction to Math course which all students take prior to taking Math I. Math I is an EOC course and counts for the schoolwide growth score while the introductory course counts as an elective for students.

The Local Condition Score was fairly close to the district average at -2, however the State Condition Score was -11 below the district average. Charles had the lowest Local Effectiveness Score at -2.5, which was still fairly close to the school system average. However, despite negative reported State Conditions, Charles fared much better than average on the State Effectiveness with a score of 9.3. Charles is a school with average Local Conditions, lower than average Local Effectiveness, high State Effectiveness, but low reported State Conditions.

School 4: Central

Central is a traditional high school serving just over 1,100 students and employing 64 teachers. The student body is the least diverse of all school in this study and consists of 86% white students and 14% non-white. At 39%, Central has a similar level of students participating in FARPL as Riley. Additionally, 92% of Central teachers are fully certified, 30% have advanced degrees, 38% are National Board Certified, and 66% have over 10 years of teaching experience. The turnover rate was 17% and class sizes are close to the school system's average.

Central did not view either local or state evaluation conditions favorably, with a score of -9.6 on the Local Condition Score and a -22.5 on the State Condition Score. These were by far the lowest condition scores in the study. Interestingly, the teachers at Central have fared pretty well on Educator Effectiveness with a local score of 1 above the district average. However, with a State Effectiveness Score of -4.2, Central had the lowest score aside from Phoenix. Overall, Central serves as an example of a school with low reported Local and State Conditions but average Local and State Effectiveness.

CHAPTER 5: Overall Trends Across All Teachers

This chapter explores trends in data across the entire sample of teachers from the study school system by analyzing responses from a survey of Math and English teachers across the four focal school sites (N=45) as well as examining the results of analysis from the focal interviews across all sites (n=14). An examination of the entire sample of responses allows for an analysis of the perceptions of a general sample of teachers to discern how evaluation policy may be related to practice, as well as to test the two assumptions of evaluation policy: (1) evaluations are necessary because teachers need to be rated, sanctioned, or rewarded in order to be motivated to do a better job and (2) evaluations yield information that is useful for teachers to improve practice.

Overall, trends across the sample of teachers surveyed and interviewed for this study demonstrate that teachers do not perceive that their evaluations provide motivation or useful feedback for improving practice. While teachers expressed a positive view of their work expectations, views about the consistency and quality of feedback from observations were less positive, and state testing data was viewed very negatively. The complementary question set from the survey showed that teachers held generally negative opinions about evaluation from the previous year with slightly more positive responses on eight of the 11 questions when anticipating the current year. Four areas on the complementary question set had statistically significant, positive changes when comparing the prior year to the current year: modifying practice from evaluation, choosing teaching strategies based on evaluation, using observation data to modify practice, and feeling evaluation will be conducted fairly. In the focal interviews, teachers stated that feedback from both the observation and testing components of formal evaluation were not useful. Teachers expressed the following concerns regarding the validity of observations: being told low rankings were necessary to show growth over years, the timing and timeliness of evaluation administration and feedback reception, the small sample of teaching actually observed, very broad or very narrow standards, unobtainable levels of distinction, and the consistency of scores across sites and administrators. The testing component was also criticized as not being timely or specific enough to provide valuable feedback. The validity of the testing component was questioned by teachers as being based on: a model that was difficult to understand, extremely low cut scores, and a small sample of both students and the curriculum.

First, I analyzed survey data to gauge the sample's overall perceptions of evaluation conditions with questions that replicated the North Carolina Teacher Working Condition Survey along with a complementary question set that asked teachers to reflect on the previous year as well as anticipate the upcoming year. Next, I analyzed interview data from 14 focal teachers across the four school sites utilizing the literature-based framework developed in Chapter 3 to explore: teacher perceptions of feedback from both elements of formal evaluation (observation and student testing), evaluation as a mechanism to motivate, reported changes in teacher practice, and teacher reform typologies. This chapter is meant to provide an overview of results from the entire sample of teachers surveyed across four schools. Chapter 6 explores how the context of the school and school-level factors may influence such perceptions and answers the research question: What, if any, role do reported school evaluation conditions and school evaluation status play in shaping teacher motivation, experiences with feedback, and work decisions related to teacher evaluation? Chapter 7 examines how the context of the individual may similarly influence perceptions and answers the research question: What individual-teacher level factors are associated with differences in teacher motivation, experiences with feedback, and work decisions related to teacher evaluation?

Survey Participants

A survey was administered to Math and English teachers at the four focal high schools in October 2016. The survey was available online through Qualtrics. The first section of the survey asked participants to provide demographic information. Table 8 outlines the demographic information for survey respondents. I have included both licensure status and years taught divided into the categories of "seven or fewer" and "eight or more." In North Carolina, a teacher is usually able to move from provisional to professional status after three years of teaching. However, due to the tenure law enacted in 2013, teachers with less than seven years of experience are subjected to full evaluation cycles every year. Therefore, it seemed pertinent to record both groups as all provisionally licensed teachers are evaluated in a full cycle, but many professionally licensed teachers are evaluated in full cycles as well.

Table 8

Survey Respondents

	Total Teachers	Taught ≤ 7 years	Taught 8+ years	Prof. License	Prov. License	Have taught EOC	English Teachers	Math Teachers	Response Rate
Riley	15	0	15	11	4	15	5	10	68.18%
Phoenix	7	3	4	4	3	7	3	4	100%
Central	13	3	10	11	2	9	7	6	76.47%
Charles	10	3	7	8	2	4	4	6	76.92%
Total	**45**	**9**	**36**	**34**	**11**	**35**	**19**	**26**	**76.27%**

Comparing Sample Teacher and School Wide Perceptions of Evaluation Conditions

In Chapter 4, I describe how publicly available statewide Teacher Working Conditions Survey data from 2016 was used to calculate school-level Evaluation Condition Scores. Table 9 shows a summary of the results of replication questions where lower numbers represent disagreement and higher numbers represent agreement. The primary purpose of asking the replication questions was to ensure that focal teachers selected for interviews did not hold beliefs that varied wildly from the average of teachers at the school. I compared focal teacher replication responses to the school averages on the replication questions as well as on the original 2016 data to determine that I was not selecting a focal teacher who held outlier beliefs.

Table 9

Responses on Teacher Working Conditions Replication Questions

Question	N	Min	Max	Mean
Teachers are held to high professional standards for delivering instruction.	44	2	4	3.57
Teacher performance is assessed objectively.	40	1	4	3.08
Teachers receive feedback that can help them improve teaching.	41	1	4	2.78
The procedures for teacher evaluation are consistent.	40	1	4	2.85
Local assessment data are available in time to impact instructional practices.	39	1	4	2.67
Teachers use assessment data to inform their instruction.	42	2	4	3.12
State assessment data are available in time to impact instructional practices.	38	1	4	2.03
State assessments provide schools with data that can help improve teaching.	41	1	4	2.22
State assessments accurately gauge students' understanding of standards.	42	1	3	1.90

Note. 1- Strongly Disagree 2- Disagree 3- Agree 4- Strongly Agree
There was an option for "Do Not Know." These responses were removed in order to calculate the means.

On average, teachers agreed with statements that were related to their quality of work overall, namely: teachers are held to high standards, teachers are assessed objectively, and teachers use assessment to modify instruction. However, there was less agreement with statements that reflect the perceived usefulness of local level evaluation to teachers when asked about the quality of feedback, the consistency of evaluation, and the timeliness of local data in order to improve instruction. When asked specifically about state testing data, teachers expressed perceptions that viewed such data in a more negative light than local data. In particular, on average, teachers disagreed that state assessments are available on time, provide feedback to improve teaching, or accurately gauge student understanding. Overall, the trends observed with the sample of English and Math teachers from the four focal schools aligned with the same trends that were observed in the district.

Teacher Perceptions of Last Year versus the Current Year about the Evaluation Process

Table 10 groups the data from the complementary question set of the survey into themes and provides descriptives and paired t-test results from the survey. The final question of the prior year section did not have a complementary current year question; therefore, a paired analysis could not be conducted for that question. In this case, the descriptives for that question are provided. Also, the paired t-tests were run question by question, excluding individuals who did not answer the complementary pair. So, there is some variability in the "n" from question to question which results from the inclusion of first year teachers who had no prior year experience to reflect on or from individuals skipping questions.

When reflecting on the previous year, each of the five of levels on the scale were used by at least one teacher for all questions. However, on average teachers seemed to disagree with nearly all of the statements. Two statements fell on average between "strongly disagree" and "disagree" and those were statements about teachers' concerns that evaluation could impact employment or label an individual as a bad teacher. All the questions about evaluation's impact on practice led to responses on average between "disagree" and "neither agree nor disagree."

Only one statement generated a response in the affirmative range and that referenced teachers' perceptions of whether the evaluation was fair.

For all but three themes, teachers' responses were higher for the same question themes about the current year compared to the previous year. All five ratings were used for each statement except for the statement, "I feel I will be evaluated fairly in the upcoming school year," which utilized between "neither agree nor disagree" to "strongly agree" and averaged as the highest overall ranking. As mentioned previously, evaluation fairness was the only theme to have a mean in the affirmative range for the prior year and had a statistically significant difference between the prior year and the current year $t(40) = -2.01$, $p = 0.05$. This significance suggests that teachers overall may have felt more optimistic about the fairness of evaluation in the current year regardless of their experiences the prior year.

Three practice related statements ranked between "neither agree nor disagree" and "agree" with means close to a neutral score of "3," demonstrating that teachers were not overwhelmingly in agreement that evaluation impacted their practice in the stated manner. However, statistically significant differences were found when comparing the themes of modifying practice using feedback from evaluation from the prior year to the current year, $t(41) = -1.83$, $p = 0.08$, choosing teaching strategies based on what one was evaluated on from the prior to the current year, $t(41) = -1.81$, $p = 0.08$, and using observation data to modify classroom practice from the prior to the current year, $t(41) = -1.83$ $p = 0.08$. Such differences demonstrate that teachers on average may have intended to more deliberately take these actions in the current school year as opposed to the previous year.

All other statements in the complementary question set fell below the neutral ranking and were not statistically significant. As with the question set focused on the previous year, the statements about concern over future employment (M= 1.93, SD= 1.07) and being labeled a bad teacher (M= 1.88, SD= 0.92) had the lowest averages though there was an overall rise from the prior year. This rise in averages between teacher perceptions from the past year to the current year suggests that teachers overall may have a more favorable outlook on evaluation in the upcoming year as opposed to the prior. However, the results do not seem to indicate that the surveyed teachers perceive that evaluations have a large impact on practices.

Table 10

Paired T-Tests of Statement Themes Reflecting on Last Year Versus This Year

	Prior		Current						
	M	SD	M	SD	n	95% CI for Mean Diff	r	t	df
Modify practice in anticipation of an evaluation	2.62	1.23	2.69	1.26	42	-0.39, 0.25	0.66***	-0.45	41
Modify practice using feedback from evaluation	2.82	1.22	3.14	1.07	42	-0.70, 0.04	0.47***	-1.83*	41
Have concern evaluation affects employment	1.81	1.11	1.93	1.07	42	-0.29, 0.05	0.87***	-1.40	41
Have concern evaluation labels as a bad teacher	1.86	1.05	1.88	0.92	42	-0.22, 0.18	0.79***	-0.24	41
Have concern evaluation does not reflect competency	2.55	1.21	2.40	1.06	42	-0.15, 0.43	0.68***	1.00	41
Choose curriculum based on what evaluated on	2.54	1.33	2.46	1.19	41	-0.27, 0.42	0.63***	0.43	40
Choose teaching strategies based on what evaluated on	2.71	1.26	2.98	1.26	42	-0.55, 0.03	0.72***	-1.81*	41
Direct focus on certain students based on what evaluated on	2.48	1.19	2.45	1.15	42	-0.26, 0.31	0.69***	1.67	41
Use test data to modify classroom practice	3.24	1.14	3.33	1.10	42	-0.42, 0.23	0.56***	-0.59	41
Use observation data to modify classroom practice	2.81	1.17	3.14	1.20	42	-0.70, 0.04	0.50***	-1.83*	41
Feel evaluated fairly	3.83	1.10	4.15	0.57	41	-0.64, 0.00	0.37**	-2.01**	40
Feel last year will impact current year	2.50	1.22			42				

Note. Scale for Survey: Strongly Disagree 2- Disagree 3- Neither Agree nor Disagree 4- Agree 5- Strongly Agree
* = p< 0.1, **= p< 0.05, *** = p < 0.01

Interview Participants

Fourteen teachers were chosen as focal participants from the four schools. Table 11 summarizes the characteristics of the focal participants.

Table 11

Interview Participants

	Total	Taught ≤ 7 years	Taught 8+ years	Prof. License	Prov. License	Have taught EOC	English	Math
Riley	2	0	2	2	0	2	2	0
Phoenix	3	2	1	1	2	3	2	1
Central	5	2	3	5	0	5	3	2
Charles	4	2	2	3	1	2	2	2
Total	**14**	**6**	**8**	**11**	**3**	**12**	**9**	**5**

After coding was complete, I first examined the frequency and the percentage of interviewees in which each code occurred. In this manner, I was able to determine that some perspectives came up much more frequently across interviewees as opposed to others (Table 12). For instance, teachers were more likely to have a negative opinion of, or to be critical of, feedback received from either component of evaluation (observation and testing). The next section discusses the trends that emerged in the interview portion along with examples and possible explanations for the trends observed. The discussion of these trends is laid out to mirror the components of the framework outlined in Chapter 3. First, I analyzed the interview data through the lens of the two assumptions of teacher evaluation policy to determine if teachers found evaluations to be motivating and/or to provide useful feedback. Next, I examined the data to determine if teachers exhibited any responses similar to those recorded in literature about teacher responses to other external reform initiatives. In doing this, I also considered whether teachers exhibited certain reform typologies during the interviews

Table 12

Interviewee Code Use

Code	Frequency	Percentage
Motivation		
Internal	13	92.9%
External	9	64.3%
Observation Feedback	14	100.0%
Negative	14	100.0%
Positive	8	57.1%
Testing Feedback	13	92.9%
Negative	13	92.9%
Positive	2	14.3%
Job Loss	8	57.1%
Work Decisions		
Strategy/How Taught	6	42.9%
Curriculum/What Taught	6	42.9%
Who is Taught	2	14.3%
Response to Reform Typology		
Acquiescence	7	50.0%
Adaptation	8	57.1%
Denial	3	21.4%

Note. n= 14

Evaluation as a Form of Motivation

Every teacher was asked whether evaluations motivated them to improve their instruction. None of the teachers indicated that their formal evaluations, either observations or test scores, motivated improvements in instruction. However, teachers noted that they were upset when parts of evaluation went poorly or if lower than expected ratings had been received. For instance, Mrs. Ranier, who had students misbehaving during an observation immediately prior to our interview, joked that she would need to drink after an evaluation like the one she had that day, “[W]hen [evaluations] fall short of what I want them to be, either justifiably, like this one, or not justifiably, like ones in the past, I just have to try to put it in another compartment of my brain, because it's so demoralizing." The disappointment Mrs. Ranier and other teachers described seemed linked to intrinsic motivation because teachers spoke of the feeling directed inward toward themselves rather than outward towards the external individual who assigned the ranking. When disappointed by their own performance, teachers may feel as though they are lacking in competency, whereas being disappointed in an unfair rating reflects frustrations with efficacy and an inability to obtain the score an individual feels the may deserve. Those who are intrinsically motivated feel good when told they are doing well and may find frustration when they either do not feel they did well or feel unjustly labelled as such.

All except one teacher offered intrinsic reasons for why they were teachers and acknowledged their own feelings of accountability as a source of motivation. Intrinsically motivated teachers were skeptical that most teachers were motivated by external factors and referenced low pay, long work hours, and lack of respect that came with their job, arguing that such conditions were at odds with someone who would be motivated by external rewards. In Mr. Allen's words, "I want to make sure that I am doing it the right way because I think it is important…I think that is really what was driving me, is that I want to make sure I am doing this because it has long-lasting impacts on these kids and on our community." Mr. Allen, like other teachers, also brought up the discrepancy he felt between his own impressions of the quality of his work and the ratings he received in observations,

> I am very aware there are some days where I go in, I am like, "Man, that was a two out of ten kind of day." It wasn't good enough. But it's for me, it is harsher coming from me than it is from a third party. Because, I do not know what their standards necessarily are. Because, I have had days I thought were very mundane days and I have gotten really

> good ratings, I am like, “No. That was not a four out of five kind of day. That was a 2.5 out of five at best kind of day.”

Nine teachers did reference extrinsic motivation by acknowledging that some teachers may need an external push, like the threat of job loss due to poor evaluations. Despite recognizing that some individuals may need to be evaluated to be motivated, these teachers did not feel that the evaluation system was particularly motivating for the majority of the teaching population. For instance, when asked about evaluation as an external motivating factor, Mr. Donaldson, an English teacher at Riley, felt that teachers would like recognition wherever it came from and in whatever form it came in, but that the system in which teachers achieved ratings was perceived as so arbitrary that most educators do not take evaluations seriously. Further pushing against the idea that teachers were externally motivated to improve work, four teachers noted that they felt fortunate that their spouses had good paying jobs that allowed them to teach and do something they really cared about as they would otherwise be unable to afford to remain in the profession.

Teachers also mentioned that when the current evaluation system was initiated it was intended that bonuses would soon be attached to their scores. A bonus policy never came to fruition, but the teachers who taught during the implementation of the current system remembered the initial plan and mentioned how this proposal resulted in increased anxiety and attention to the evaluations initially. However, concerns were alleviated when the bonus system never materialized. Overwhelmingly, the focal teachers interviewed for this study did not view evaluation as a means to motivate individuals to do better at their job and cited intrinsic sources of motivation as more valuable motivators for the improvement of practice.

Evaluation as a Source of Feedback

Feedback from evaluations was the topic that dominated conversations about evaluation. In the formal evaluation policy examined in this dissertation, feedback stemmed from both administrator observation and state testing. Both sources were referenced nearly equally across interviewees. In general, teachers usually critiqued the feedback that was provided by both components of formal evaluation with all focal teachers referencing negative aspects of observation feedback and all but one for testing feedback.

Specifically, teachers critiqued formal observations due to systematic concerns about the growth model promoted by the state and the small sample of observations conducted. Other critiques included concerns related to: the timing of both when observations were conducted and when feedback was received, some standards being either too broad or too specific, the difficulty of achieving high marks, and the lack of standardization and consistency in scoring across sites and evaluators.

Regarding testing, teachers had difficulty with understanding the way in which their growth scores were calculated. Teachers explained that the metric used had either been not explained to them or was difficult to understand. Teachers also identified mathematical weaknesses in their understanding of the model. For instance, teachers referenced that the way in which the teacher's growth score was calculated did not seem to truly account for the small sample size of either students or questions used to calculate scores. Additionally, teachers raised concerns with student issues that fell outside of the teacher's control (such as frequent absences or extended illnesses) yet impacted a teacher's growth score nonetheless. Teachers also questioned the validity of the test used to calculate student growth and described incredibly low cut scores which allowed students to pass or even obtain high grades with low percentages of correct questions. Finally, teachers stated that they were unable to utilize the feedback provided by testing because of the amount of time it took to receive, and the lack of specific information provided.

Perceptions of Feedback from Observation

The timeliness of evaluations and the amount of time spent on evaluations were problematic for teachers who referenced evaluations being done at inopportune times or in sequences that were unable to best assess teaching. For instance, teachers on full evaluation cycles often brought up instances where all three required evaluations would occur in quick succession within a month, often in the same class, rather than sampling throughout different classes during the year. Teachers indicated that they felt this sampling technique made it impossible for an administrator to really gauge how a teacher handled different types of classes and different groups of students. Teachers who were observed exclusively in a difficult class felt at a disadvantage to those who may have been exclusively observed in a higher achieving or better-behaved class.

Additionally, teachers stated that this quick succession of observations often occurred later in the year when administrators expressed that they were trying to make up for missed observations and complete requirements before a deadline. The teachers who experienced this lamented being observed during review time for exams when they were unable to demonstrate how new concepts were taught to students or were required to participate in certain review activities instead of "actual teaching." An additional criticism teachers raised was having an observation prior to receiving feedback from an earlier observation, which prevented teachers from learning from the first observation and addressing issues that may have been brought up, or ensuring the next observed lesson exhibited standards that may not have been met in the previous observation.

Teachers also expressed frustration with perceived flaws in the structure of the observation system itself. Teachers at all four schools noted that when the current evaluation instrument was new, when they were early in their career, and/or when a new administrator evaluated them for the first time, they were told by the observer that they would be rated lower initially to allow the teacher to “show growth” in future evaluations. Mrs. Ranier, a veteran English teacher at Charles explained, “I have talked to administrators and I know they were told when they were trained to evaluate us that they have to leave room for growth. Which means you cannot ever be at the top, not really.” The approach described by Mrs. Ranier was brought up by teachers at all four focal schools. Another English teacher at Charles, Mr. Eagle, who had previous teaching experience but was new at the current school, was told prior to his first observation of the year, “You are going to be developed or proficient, the very first two categories. You will not hit advanced, there is just no way you are going to hit distinguished.” This left teachers with a sense that the scores received from an observation may not be a true, objective reflection of teaching ability and instead a score the administrator gave that was subjective to how long an administrator had observed them and would allow room for the administrator to show that a teacher has “grown” over time.

Additionally, several teachers argued that an observation of a fraction of a class period, even if conducted three times a year, did not provide an adequate sample for an administrator to get an idea of a teacher’s ability. Mr. Allen, an English teacher at Central, began to calculate the actual amount of teaching his administrator observed as compared to the amount of time he taught the entire year, explained, “If a statistician looked at that, they would be horrified that is how we get evaluated." Teachers stated that it would be preferable if administrators took more interest in the work of teachers outside of evaluations so that observations were not the only time the evaluator was exposed to an individual's teaching practice.

Administration taking a greater interest in teacher work could manifest in a few ways. Two of the schools, Community and Riley, required teachers to submit daily lesson plans, but teachers also felt that having engaging pre- and post-conferences (a requirement of the evaluation system for those undergoing full evaluation cycles, which was reportedly not followed with fidelity) and having administrators pop into classrooms more frequently for informal check-ins would lead to a better assessment of teaching, rather than a few formalized snapshots each year. Mrs. Ranier, who had taught for 22 years, suggested that instead of taking large chunks of time for formal observing followed by lengthy post conferences, principals should be more present in classrooms in order to better know the staff and their teaching. Reflecting on schools where she had previously worked, this teacher stated,

> I really enjoy working at a school where the administrators are present and they are often in your class, because then when I sit down with them, and they are talking about my teaching, they can say, "Oh, but you did this on the other day, when I was just walking through." And also, then their presence, in and of itself, would not be so disarming when they come here the two or three times they come to do an actual observation…they should be more present, so that I feel more comfortable, and so the students feel more comfortable with them, and they have a better idea of whether or not I am doing my job consistently.

This critique was common across interviews and while I did not specifically ask each teacher about how often administrators visited informally, no teacher reported that an administrator was ever in their classroom in the study year aside from formal observations.

Teachers also talked about certain evaluation standards being either too specific or too broad, which they perceived made it difficult to receive feedback that could actually be used to help instruction. One of the standards that was mentioned frequently was the technology standard. Some teachers had difficulty meeting this standard because their administrators only showed up to observe at times when technology was not used. Mr. Allen discussed the difficulty he had satisfying the technology standard, “One of the standards is: Do you use technology? If I am seen four times a year…I have had years where I get ‘no’ or I will get just whatever the bare minimum one is... I use technology most weeks, two or three times a week. [The administrator] came in four times and it was four days where in that particular period I was not using technology.”

Conversely, some Math teachers brought up how easily satisfied the technology requirement was because their observing administrator had a poor definition of what constituted technology. Evidently, some observing administrators counted calculator use as technology use to satisfy the requirement of the observation rubric. Therefore, Math teachers would not need to use any other technology source to meet that requirement of the evaluation other than a tool which was already commonly used and required as part of the curriculum.

However, every Math teacher interviewed referenced difficulty meeting another standard on the evaluation rubric. Each of the Math teachers brought up how the standard “global awareness” presented a challenge. I asked one teacher, Mrs. Proffitt to describe what global awareness looked like in a Math classroom and to explain what was meant by how resources given to Math teachers to satisfy the requirement were lacking. She described attending a professional development training for a program called “Newszilla” which teachers were told they were expected to use and could accommodate the global awareness standard on the observation rubric, “I looked and looked and there were like two that actually applied…It is not the same in every classroom. Social studies, English. Science, you can bring it all in, but it is very difficult for us in Math to.” Another Math teacher, Mr. Robbins, summed up the issue this way,

> I feel I am at a disadvantage compared to a History teacher. Even a Science teacher, I think, would have a little bit easier time with that because when you talk about different events like pollution… you can talk about what is going on in different countries. Mathematics, if we are studying quadratic functions, how am I supposed to incorporate global awareness into that without using some kind of a stretch of a real-world context that is just really bizarre?

Additionally, two Math teachers noted that the stretch that was necessary to incorporate the requirements of the “global awareness standard” fell outside of and may even have opposed the Math standards set by the state. These teachers contended that the standard and the push to use programs such as “Newszilla” required teachers to teach things that were outside of the established standards and outside of what was tested. So, these Math teachers felt that the requirements of the teacher evaluation process were at odds with the actual standards of the courses taught.

Conversely, teachers also pointed out that some standards seemed far too broad. This issue was raised by others who mainly indicated that the broad interpretations of standards in the observation rubric did little to promote the more standardized observations promised when the current evaluation instrument was introduced. Mr. Robbins explained how a standard about teachers being ethical felt too broad to be a standard that should be observed, “And I was always bothered by that because I felt like I should be in the distinguished, the farthest up possible. But the only way that I have been told that I think I could even get to that is by basically doing a workshop for teachers teaching them the code of ethics.” Mr. Robbins felt that not being ranked high in the ethics category indicated a deficit in ethics, whereas the rubrics required the sharing of knowledge amongst staff through activities such as workshops or leading staff staffing to be a pre-requisite for achieving high marks.

Several teachers, including Mr. Robbins, were troubled by the absence of standards that focused distinctly on a teacher’s ability to teach the subject area. Mr. Robbins explained, “There is nothing in those standards that says anything about Math.... It is just, ‘Are you a good teacher?’ It is very broad.” This was a topic broached by teachers of both subject areas; however, the relationship between observation and subject area will be explored in greater detail in a subsequent chapter on individual-level differences. For the conversations with focal teachers, it appears that the assessment of whether a teacher was competent in their respective subject area was left up to the measurement of student growth as calculated by standardized testing rather than by an administrator’s judgement.

Similarly, while not necessarily referencing specific standards, the difficulty of obtaining high rating levels such as “distinguished” was brought up by other teachers. Mrs. Ranier described how it would be impossible to be considered a distinguished teacher overall, even after 22 years in the classroom, based on the observation rubric used,

> I do not like the idea that I have to do things outside of the classroom that I do not want to do… I do not want to go to professional meetings, I do not want to lead a committee, I do not want to do any of that… So, I do not like the new evaluations in that I am graded

> for things that I do not feel are the reasons I became a teacher, and reflect my performance and abilities as a teacher... I help to hire new teachers? No. Am I part of a professional organization? No.

Overall, teachers expressed that they felt it was nearly impossible to be highly rated as a teacher overall due to the requirements of the protocol.

Teachers also indicated that they felt the observations and ratings were not standardized or consistent. However, achieving higher ratings may be easier in in locations outside those in this study. Four teachers referenced looking at the publicly available scores at other schools and noticing that in some locations, all teachers were rated in the highest two designations for all categories. Critiques of such practice was one of the reasons used to justify the current, lengthy evaluation protocol used in North Carolina and teachers were quick to point out that the issue had not yet been remedied (Weisberg et al., 2009 and others). Teachers also mentioned discrepancies in evaluations between different schools in which they had worked or even within the same school between evaluators. Teachers indicated that they knew which administrators in the school would be "tougher" on observations than others. Teachers also seemed aware of which administrators would provide better feedback and which were just "checking a box." Most of these issues are related to the context of site or of the individual and are discussed in the next two chapters. However, it is important to note that teachers overall expressed that they did not feel that they could accurately trust their ratings as a reliable source of feedback about their practice due to discrepancies in how these rankings were awarded.

Overall, teachers were very critical of the feedback received from observations and of the observation process as a whole. Nine teachers mentioned a belief that the evaluation process as a whole may be necessary to help expedite the formal removal of teachers who should not be in the classroom, but all felt the current system used was grossly ineffective at providing feedback to improve practice. Eight teachers overall mentioned positive aspects of the formal observation process and described how it created an opportunity for them to independently reflect on their own practice. However, these same teachers stated that reflective practice was something they already did on a regular basis and the observation was just an opportunity for them to engage in such behavior more systematically.

Perceptions of Feedback from Testing

Similar to observations, teachers also expressed frustration with the feedback received from the testing component of their evaluations. There are two types of tests administered to high school students that count towards teacher evaluation scores: the End of Course (EOC) exam which is given in Math I, English II, and Biology and counts for both individual teachers and for school-level evaluations; and the North Carolina Final Exam (NCFE) which counts for individual-level teachers and is administered to all other courses, with few exceptions. In the case of this study, those exceptions include Advanced Placement courses and in the case of Charles High School, an Introduction to Math course which counted as an elective that the school required prior to Math I.

There were two concerns which were unique to subject area. First, English teachers expressed frustration that the tests covered a small amount of the standards for their subject area. Second, Math teachers expressed concerns over recent changes to the curriculum and tests. These issues will be explored deeper in Chapter 7 which focuses on individual-level characteristics like subject area.

Both Math and English teachers expressed concerns over the accuracy of the test, particularly in regard to the scoring scale that was used and the method used to calculate standard

6. Teacher growth for to standard 6 is calculated using a very complicated psychometric model and teachers seemed to know little about it aside from that it was supposed to calculate how much a student grew in a given year and that the highest and lowest scores were eliminated as outliers. Teachers expressed frustration that the score was calculated with such a small sample of questions (40) and two teachers cited examples of outlier scores which were dropped despite the teacher feeling that the student had made significant amounts of growth due to the work of that teacher. Teachers also believed that the calculating technique was ineffective at eliminating poor results due to other factors outside teaching, such as in the case of excessive student absences or illness during test day.

Aside from the two subject area-specific concerns referenced earlier, teachers also questioned the validity of the test in regard to the cut scores used for students. One teacher explained, “One student got, out of 40 questions, she got seven right, and that was not a pass, but it was still very high, it was like a 58, and I remember thinking, ‘I should have just told her to pick ‘A’ for every answer because she would have done better than getting seven right. Similarly, a math teacher quipped, “Well, if you are curving it down that far, if you are guessing on every single one, the difference between an A and a C [for a grade] is negligible.” The cut scores in particular seemed to frustrate teachers as these seemed to provide students with an unreliable indicator of their performance that did not mirror the scale of assessments used by teachers in the classroom.

Another major barrier to using the tests as source of feedback was the timing and specificity of the feedback received. While the raw score and the grade of the student was received quickly following the test, the breakdown of scores was not received until the following fall when the new school year was already well underway. Once the more detailed reports are received the following fall, teachers expressed further frustration over the level of detail provided as particular goals are not identified in the data. An English teacher explained, "[It] is always frustrating to us that the data we receive back from the test itself is just so general. It will say 'Reading Information Strand 2,' and Strand 2 has six different goals in it, so I have no idea where my weaknesses truly are as a teacher, to help the kids grow there." Teachers also perceived that it was impossible to discern from the data whether the scores were a result of instructional decision of the teacher. One teacher explained how his students exhibited a higher than usual amount of growth in the last testing cycle, "I do not know where that happened. I know of different things I did, but I do not know if that actually had made that [difference]. I do not know if my changes made the growth happen or is it just because maybe we read some stories that related to the story that was on the test that year and they were just more familiar in some way." Overall, teachers perceived a lack of specific, actionable feedback as a barrier to making improvements.

There were only two teachers who felt that the feedback they received from testing was positive and both taught at the alternative school, Phoenix. In both cases the teachers referenced historical data rather than the data received in a testing cycle completed by that teacher. Mr. Forest, the Math teacher at Phoenix, told an anecdote about looking back on past test data for a student and realizing he had missed Math courses,

> We looked at his test data… and he was fairly consistent in elementary school, he was probably scoring threes and fours on the state exam tests, and then in fourth grade, it just tanks, and he's down in level 1, level 2…We kept digging and digging, and we looked into his transcript, and he had not actually taken seventh and eighth grade math. And then now, he is 18 and he is trying to learn Math 1…He is bad at fractions, that basic

> background, even to have to be able to just kind of push him through Math 1, 2, and 3…
>
> So, that kind of addresses how we can use the state data.

So, in the above example, the record keeping that follows state testing was useful in providing a teacher feedback on what was missing from the student's background. However, other than these two teachers from the alternative school who offered similar examples, the other teachers in the sample did not describe having any positive experiences with regard to the feedback received from testing data.

Feedback from Other Forms of Evaluation

The literature reviewed in Chapter 3 illustrated that feedback can be beneficial to improving a teacher's performance; however, Hackman and Oldham (1980) argued that feedback should stem from the work of teachers. When discussing the formal evaluation system utilized in North Carolina, teachers seemed to be referencing a disconnect between their work and the feedback provided. Rather than being driven by the work, feedback often seemed to be driven by the observation instrument or the values the observing administrator had derived from the instrument. Similarly, the tests used to calculate standard 6 of teachers' evaluations presented feedback that was driven by the test itself as well as by the values driving the test.

Interestingly, when asked about the relationship between evaluation and feedback, the focal teachers often referenced other types of evaluation and sources of data outside of the formal evaluation process. These references usually entered the conversation organically without any prompting. However, once this pattern was discovered, I started asking teachers about their experiences with other forms of evaluation, whether that be observation or testing, when they had not mentioned other sources of feedback.

Eight teachers referenced experiences with other observers, such as instructional coaches and in one case, the superintendent, to be beneficial and meaningful. Mr. Robbins explained the relationship he has with his instructional coach,

> We have a Math coach here. She comes at least once a week, sometimes twice a week, and will come in and observe our classes. She offers assistance during that class. Occasionally, she will help students or I can ask her a question about, “Where do I go now in this lesson?” And she can direct me. Sometimes she just comes and observes, and then later, she will come during my planning period and talk to me about what she saw, what she noticed… She will come back another week and do it again, and maybe say, “Hey, I noticed that you tried this today and it worked really well.” So, I am getting evaluated from her, but it's not on any kind of formal basis.

Mr. Robbins goes on to explain how this type of observation counters many of the complaints about the feedback from formal evaluation described previously. He explained that his instructional coach observations occurred much more frequently than formal observations and highlighted the coach’s background and expertise in Math as being of particular importance and relevant to the quality of feedback that he received from observation experiences.

The experience of Mr. Robbins echoed that of many other teachers who brought up informal evaluation experiences as more valuable than the formal administrator evaluation required by law. Such teachers highlighted the frequency of the observations, the personalization of the experience, and the subject area knowledge of the observer as key components of what made them feel that these informal evaluation experiences more successful. So, teachers are not dismissive of all forms of observation feedback and many in this study found other, informal sources as useful. Rather, teachers presented legitimate concerns over the usefulness of the feedback received from formal evaluations which were often countered by more positive, informal experiences.

However, not all teachers indicated that informal evaluations were positive. The value of the experiences seemed to be contingent on the people involved and the respective teaching situation. Four teachers felt like the coaches brought in theory that was not applicable to the realities of the classroom the teacher was working in. One Math teacher explained, “There is theory and then there is practice…I have not found [instructional coach observations] to be useful to me because it is like asking me to really reconstruct how I am going to teach and I am not going to be able to do that. And so, the ideas that I am getting are not really ideas that I can implement that are realistic for me to try.” In this case, the informal observer may have failed at relating feedback to the teacher’s work as her approach was more theory-driven than based in practical application.

As with observation, teachers also mentioned other methods of testing which were informal and provided feedback that was useful to their practice. Specifically, Math teachers mentioned the use of county-wide benchmarks. The focal English teachers did not use county benchmarks, though they were aware of them, and some veteran teachers mentioned using them in the past. However, teachers of both subjects mentioned using school-based common assessments designed in professional learning communities (PLCs). One English teacher described how schoolwide common assessment worked at Central, “We design the pretest, we teach our unit, and then we give our post-test and then compare those to pre and post-test assessments, but then we sit down and we break them down into smaller components in order to then revise our instruction.” She described this experience as “authentic” and explained how this type of assessment design provided feedback that benefitted her instruction and allowed teachers to more readily evaluate why a student may have missed a question, “I know exactly which questions my kids missed, so I know where my weaknesses were in my instruction. So yes, I use that all the time to inform my instruction… and I can also look at that particular question and say, ‘Was it how the question was worded or was it the skill?’”

Again, the interviews suggest that teachers are not totally dismissive of testing as a form of feedback. Instead, teachers stressed that the informal tests which were successful were designed to match closely with the work of teachers and to provide feedback in a manner that could be used by the teachers. The informal testing described here provided feedback that was both timely and specific enough to show teachers areas of strength and weakness that could be improved upon with their current set of students.

Responses to Reform

There is some evidence that teachers change practice either in anticipation of or resulting from teacher evaluation policy in manners similar to those observed in teacher responses to other external accountability pressures. Overall, nine of the 14 interviewees mentioned changes in practice due to either aspect of the evaluations (observation or testing). Six focal teachers referenced changing teaching strategies or how a teacher taught, six focal teachers described changing curriculum or what a teacher taught, and two referenced focusing on certain students due to evaluation. The way in which these responses manifested will be subsequently described.

The changes in teaching strategies that the six teachers cited were generally quite superficial. For instance, teachers mentioned making sure that learning targets were listed on the board because they knew their administrator would check for those during observation. Additionally, teachers described trying to “hit a box” on the evaluation score sheet for a standard such as technology use, which had not yet been observed.

Teachers also adapted their teaching strategies based on testing, but again this often resulted in minor alterations. For instance, one Math teacher explained how he made his students complete warm-ups on the computer to mirror the conditions under which students would take the state assessment. Similarly, another Math teacher described how the format of Math problems may be altered to better mirror what students would see on a test, “If students see an equation written a certain way, I know to make sure to show them that format versus another format that is not incorrect.” Likewise, an English teacher at Riley described creating study guides with questions that were worded in the same manner that students would find on the standardized test. All of the alterations described were superficial changes that put students in situations that better mimicked the conditions of state testing.

When asked about making changes to curriculum due to evaluation, the six teachers referred to the influence of the state tests as opposed to observations. For instance, an English teacher at Central described how she focused more on reading curriculum, which she knew would be on the test, at the expense of other, non-tested elements of the English curriculum (specifically the writing, speaking, and listening strands of the Common Core State Standards). The same teacher stated that this was especially true for her standard classes as opposed to her honors courses, with the former needing much more of a “push” in order to show growth on the exam.

Mr. Forest, the Math teacher at Phoenix, saw evaluation policy as ideally being aligned with good teaching and referenced both the observation and testing components in his explanation. He explained that he felt that choosing curriculum that aligned with the way he would be evaluated was a “moral obligation.” Additionally, he stated that he was not motivated by the evaluation itself, but instead by the principles of good teaching that the evaluation was meant to measure, “When I am looking at the standards, I see that as what a teacher should do already, and I know that most other teachers at my school feel for the most part the same way, that we do not plan for that evaluation.” Mr. Forest reiterated that he sees the evaluation standards as something he should be doing on average most of the semester and felt this approach may seem lax, but that at his particular school his job was not in danger as there was a lack of Math teachers, particularly those willing to teach at an alternative school.

There were only two interviews that referenced focusing on particular students due to the formal evaluation. These examples were about directing certain skills at students who needed increased help passing the test; for instance, teaching reading strategies to students who read below grade level or basic Math concepts to students who lacked the skills to complete grade appropriate work. There were no mentions of directing attention on certain students due to observations. Such behaviors were also very superficial and did not represent radical changes in a teacher’s practice.

Reform Typologies

Reform typologies were recorded if a teacher made a statement during an interview that demonstrated that they fit one of the three categories of Yurkofsky’s condensed reform typologies: acquiescence, adaptation, or denial (2016). Seven focal teachers included statements that indicated acquiescence. Teachers who demonstrated acquiescence generally stated that they accepted their evaluations without question. For example, Mr. Augustus, a Math teacher at Central called formal evaluation, “a fact of life.” Teachers who demonstrated acquiescence did not feel that evaluations had any effect on their teaching lives in any other way, nor did they try to push back on the system or attempt to adapt the system to be more useful for them. For instance, Mrs. Street, an English teacher at Phoenix surmised, “I do not even really pay attention to the principal being in the room…I always ask him, ‘Do I still have a job?’ And I still have a job, so it went well.”

Some teachers referenced ways in which they were able to take formal evaluations and adapt the process and the results into something that was useful to them. Overall, eight focal teachers included statements that exhibited adaptation. For example, at Riley, teachers were required to submit weekly lesson plans and the new administrator there connected these submissions to the evaluation scores given to teachers. Mrs. MacDonald, an English teacher at Riley explained how she adapted this policy to suit her needs, “Instead of having something a week out that I am going to have to spend the whole week revising anyway…I am just doing it a day at a time, and [the principal] has not said anything about it. It seems more manageable.” Other teachers described how the evaluation process was more of an opportunity to self-assess practice, which would represent another form of adaptation. One teacher described how, rather than worry about ratings from a third party, she looked at the standards as a sort of checklist against which she could rank herself.

There were two teachers who described actions or attitudes that ignored the policy or indicated denial about it. While these teachers still participated in the requirements of the evaluation, they mentioned ignoring some directives related to them. For instance, Mr. Donaldson was identified as weak in one area by an administrator and was told to go back and change his Personal Development Plan (PDP) to reflect improving that weakness as his goal for the year, he stated that he never did it and explained, “I am not going to go back in my PDP and change it because it is not something that I feel will make me a better teacher.”

Denial of the evaluation policy was a sentiment echoed by Mr. Forest at Phoenix who questioned the Math competency of his administrator and stated,

> I am going to do what I want to work on. I am not going to work on what [my administrator] tells me to, just because [he] just told me… I believe that I should be free to pick my own things to work on. And I think, if a teacher is unwilling to do that, then the principal should then be able to step in and actually accurately say, “These are the things you should work on.” And that is what an evaluation should give you. But I think if the teacher is doing that on their own, it is public enough, we do not need to do an evaluation and rank like that.

In addition to refusing to adjust goals based on his principal’s suggestions, this teacher was particularly upset that his principal seemed to lack an understanding of what “good Math teaching” looked like. In response, he coupled with colleagues from other schools to write a grant proposal that would train principals to recognize good Math practices to help them not only with evaluation but in recruiting and retaining effective Math teachers.

Conclusion

Overall, an examination of the trends across the sample of teachers surveyed and interviewed for this study shows that teachers may make minor alterations to their practice due to evaluation policy. However, these changes are far from revolutionary and instead represent very superficial adjustments rather than deep, meaningful, and sustained changes. Moreover, the focal teachers did not self-report that formal evaluations were a motivating force, nor that feedback was useful in improving their practice.

An examination of the TWC replication questions yielded results that mirrored trends found overall in the 2016 data. Overall, teachers expressed a positive view of their work expectations, a less positive view about the consistency and quality of feedback from evaluations, and a very negative view of the value of state testing data. The complementary question set from the survey showed that teachers held generally negative opinions about evaluation from the previous year with slightly more positive responses on eight questions when anticipating the current year. However, responses still demonstrated an overall negative perception of the evaluation process. Four areas on the complementary question set had statistically significant, positive changes when comparing the prior year to the current year: modifying practice from evaluation, choosing teaching strategies based on evaluation, using observation data to modify practice, and feeling evaluation will be conducted fairly. These changes may be driven by staffing and initiative changes in some of the schools, which are explored further in Chapter 6.

In the focal interviews, teachers stated that feedback from both the observation and testing components of formal evaluation were not useful. Teachers expressed the following concerns regarding the validity of evaluations: being told low rankings were necessary to show growth over years, the timing and timeliness of evaluation and feedback reception, the small sample of teaching actually observed, a combination of very broad and very narrow standards, nearly unobtainable levels of distinction, and the consistency of scores across sites and administrators. Teachers expressed frustration that formal evaluations were not supplemented with a continuous informal presence of administrators in the classroom. While teachers overwhelmingly had negative views of observation, two interviewees mentioned that the observations provided them with an opportunity to be reflective, though these teachers stated that this was something that was ingrained in their practice anyway.

The testing component was also criticized as not being timely or specific enough to provide valuable feedback. Teachers also questioned the accuracy of the tests due to dramatic cut scores and small samples of both questions and students. Additionally, teachers had concerns that the equation used to calculate their growth as a teacher was opaque and seemed inaccurate. Two interviewees referenced archived testing data as potentially being helpful to identify past student weaknesses, though neither teacher found testing data received for current students to be particularly meaningful. Many, but not all, teachers described feedback received through other sources of informal evaluation, most notably observation by curriculum coaches and locally developed tests, as sources of feedback that were more meaningful and more useful than feedback from either component of formal evaluations.

Possibly for the reasons discussed in the feedback section, teachers did not feel that they were externally motivated by evaluations, though a few mentioned that a poor evaluation would be upsetting to them. The disappointment teachers described was framed as resulting from feeling personally let-down by poor performance (whether real or perceived) and indicated the teachers were intrinsically motivated to do well. Teachers also referenced several discouraging aspects of teaching and described intrinsic rewards of teaching as the prime motivator for doing well in their job. With this in mind, it is not surprising that teachers do not feel that evaluation has much effect on their practice. Changes that were mentioned by teachers consisted of superficial issues like listing learning targets on the board, trying to incorporate technology into an observation, or adjusting classroom activities like warm-ups to more closely match the format of the state test. Interestingly, when teachers talked about how evaluation influenced practice, they referenced both observation and testing as affecting how they teach, but testing was overwhelmingly referenced when making choices in curriculum or when directing focus on certain students. Not every teacher demonstrated a reform typology in the interviews, but acquiescence and adaptation were more common than outright denial of the policy. Given that teachers are legally bound to abide by the policy, it is not surprising that most teachers would be unwilling to totally disregard or even push back against the policy.

CHAPTER 6: The Context of the School Site

The previous chapter presented results across the entire sample population for this study. This chapter answers my first research question: what, if any, role do reported school evaluation conditions and school evaluation status play in shaping teacher motivation, experiences with feedback, and work decisions related to teacher evaluation? As described in Chapter 4, the four focal schools for the study were selected due to variability in evaluation conditions (using measures created from the 2016 administration of the North Carolina Teacher Working Conditions Survey) and evaluation scores (using 2016 Educator Effectiveness data). The hypothesis that motivated this selection was that schools where teachers perceived evaluation conditions to be very good may view the impacts of evaluation on practice differently than schools where teachers perceive conditions to be very poor. Similarly, teachers from schools with highly rated teachers (i.e., receive high evaluation scores) may perceive the impacts of evaluation on practice differently than schools where many teachers receive low scores.

The schools in this study represented a range of evaluation conditions and effectiveness scores. Riley demonstrated Local and State Evaluation Condition Scores which were very close to the district average (0.6 and -1). Phoenix demonstrated a slightly above average Local Evaluation Condition Score (4.4), while Charles had a slightly below average Local Evaluation Condition Score (-2). Central had a very low Local Evaluation Condition Score (-9.6). Low State Condition Scores were demonstrated by Charles (-11), Phoenix (-16), and Central (-22.5). In this study, all of the schools demonstrated Local Evaluation Effectiveness Scores that were very close to district average with scores ranging from -2 to 1. However, the schools demonstrated variety in State Evaluation Effectiveness Scores. Riley was near district average (1.5) while Central was slightly below (-4.2). In contrast, Phoenix was far below district average (-13) and

Charles was far above district average (9.3). So, the four schools in this study offered a wide range of varying combinations of the two scores.

In this section, I first present an analysis of the survey data to determine if there were statistically different perceptions held among school sites as measured by the survey results Then, I use a combination of survey and interview data to explain how evaluation conditions and evaluation scores are related to the perceptions of teachers at the four focal schools as well as present some alternative explanations based on the interview data. Finally, I describe the specific evaluation scenarios present in the school contexts in this study.

Comparing Perceptions of Evaluation and Practice between School Sites

As explained in previous chapters, the final section of the survey contained complementary thematic statements that asked teachers to reflect on the previous school year as well as anticipate about the current school year using a Likert scale. Table 13 shows the descriptive statistics for each complementary question set separated by school site.

Each thematic set was analyzed using a one-way ANOVA to determine if there were differences between schools in each complementary set. No significant differences between schools were found on any of the questions from the complementary questions set. However, significant differences did emerge within schools when comparing the prior to the current year, which will be discussed in an upcoming section. Later in this chapter, I examine the cases of each school and elaborate on how evaluation conditions and evaluation scores may have impacted teacher perceptions of evaluation in ways in which the survey was unable to capture.

Table 13

Complementary Question Set Means by School

		Riley			Phoenix			Central			Charles		
		M	SD	n	M	SD	n	M	SD	N	M	SD	n
Modifying practice in anticipation of an evaluation	Prior	2.47	1.41	15	2.60	0.89	5	2.67	1.30	12	2.80	1.14	10
	Current	2.53	1.51	15	2.71	1.25	7	2.83	1.15	12	2.90	1.29	10
Modifying practice using feedback from evaluation	Prior	2.73	1.28	15	2.80	1.48	5	2.50**	1.12	12	3.30	0.95	10
	Current	3.20	1.08	15	3.43	0.98	7	3.25**	1.36	12	3.00	0.94	10
Concern evaluation affects employment	Prior	2.13	1.46	15	2.20	1.64	5	1.42	0.52	12	1.60	0.52	10
	Current	2.33	1.50	15	2.43	1.27	7	1.58	0.52	12	1.70	0.48	10
Concern evaluation labels as a bad teacher	Prior	2.00	1.20	15	2.00	1.73	5	1.67	0.65	12	1.80	0.92	10
	Current	2.13	1.19	15	2.00	1.00	7	1.83	0.84	12	1.70	0.48	10
Concern evaluation does not reflect competency	Prior	2.53	1.51	15	2.80	1.64	5	2.25	0.97	12	2.80	0.79	10
	Current	2.33	1.11	15	2.71	1.25	7	2.42	1.24	12	2.50	0.71	10
Choosing curriculum based on what evaluated on	Prior	2.73	1.39	15	2.40	1.52	5	2.33	1.30	12	2.50	1.27	10
	Current	2.67	1.35	15	2.50	1.23	6	2.17	1.03	12	2.80	1.23	10
Choosing teaching strategies based on what evaluated on	Prior	2.40*	1.18	15	3.20	1.30	5	2.92	1.44	12	2.70	1.16	10
	Current	2.80*	1.27	15	3.71	0.95	6	2.83	1.47	12	3.10	1.10	10
Directing focus on certain students based on what evaluated on	Prior	2.53*	1.19	15	2.80	1.30	5	2.33	1.30	12	2.40	1.74	10
	Current	2.33*	1.23	15	2.71	0.76	7	2.25	1.22	12	2.70	1.16	10
Use test data to modify classroom practice	Prior	3.60	0.99	15	2.20*	1.10	5	3.17	1.40	12	3.30	0.82	10
	Current	3.60	0.99	15	3.29*	1.11	7	3.00	1.35	12	3.40	0.84	10
Use observation data to modify classroom practice	Prior	2.67	1.11	15	3.20	1.10	5	2.83	1.47	12	2.80	1.03	10
	Current	3.20	1.32	15	3.71	0.95	7	2.92	1.38	12	3.10	0.88	10
Feel evaluated fairly	Prior	3.60	1.55	15	4.00	0.71	5	4.17	0.58	12	3.70	0.68	10
	Current	4.07	0.70	15	4.14	0.69	7	4.36	0.51	11	4.00	0.00	10

Table 13 (cont'd)

Last year will impact current year	Prior	2.33	1.23	15	2.80	1.10	5	2.42	1.38	12	2.70	1.16	10
	Current												

Note. Scale for Survey: Strongly Disagree 2- Disagree 3- Neither Agree nor Disagree 4- Agree 5- Strongly Agree; * = p< 0.1, **= p< 0.05, *** = p < 0.01

Insights from Focal Teacher Interviews

I conducted interviews with 14 focal teachers across school sites and coded transcripts to discern differences between motivation, feedback use, reform responses, and reform typologies across schools. The data in Table 14 presents the interview case for each code. In other words, the frequency is the number of interviews that included at least one occurrence of the code. Using this reporting method rather than frequency of interviewees was necessary due to the small number of participants in this data set when separated by school and the variable number of interviewees at each location. Each interviewee participated in two interviews for the study, so the "n" reported equals twice the number of interviewees at a school. The percent of occurrence for the interview at each school is also reported to allow for a better comparison across schools with varying sample sizes.

These data will be examined more in-depth for the section on "Evaluation Scenarios" at the end of the chapter. However, the most notable differences between schools center on interviews from Phoenix. For instance, Phoenix had the highest occurrence of the reform typology of acquiescence (which occurred when a teacher made a statement that indicated an acceptance of the policy, which could be with reluctance, but without protest) of the four focal schools with 50% of the interviews containing statements that indicated acquiescence. Phoenix teachers also mentioned internal motivation more frequently by percentage of interview, at a rate of 66.7%. Conversely, at Riley there were internal motivation statements in only 25% of the interviews. Additionally, Phoenix interviews were the least likely to mention positive aspects of observation (16.7%), but it was the only school where positive aspects of testing were mentioned. It may be that the unique circumstances of the alternative school account for the differences when compared to the three traditional high schools across interviews. The context of Phoenix will be explored in the next section.

Table 14

Code Interview Case Count by School

	Riley		Phoenix		Central		Charles	
	(n= 4)		(n= 6)		(n=10)		(n= 8)	
	Frequency	Percentage	Frequency	Frequency	Frequency	Percentage	Frequency	Percentage
Motivation								
Internal	1	25.0%	4	4	4	50.0%	4	40.0%
External	2	50.0%	3	2	2	25.0%	3	30.0%
Observation Feedback	2	50.0%	7	6	6	75.0%	7	70.0%
Negative	2	50.0%	7	6	6	75.0%	7	70.0%
Positive	2	50.0%	3	3	3	37.5%	3	30.0%
Testing Feedback	2	50.0%	6	6	6	75.0%	6	60.0%
Negative	2	50.0%	6	6	6	75.0%	6	60.0%
Positive	0	0.0%	0	0	0	0.00%	0	0.00%
Work Decisions								
Strategy/How Taught	1	25.0%	1	4	4	50.0%	1	10.0%
Curriculum/What Taught	0	0.0%	2	2	2	25.0%	2	20.0%
Who is Taught	0	0.0%	1	0	0	0.00%	1	10.0%
Response to Reform								
Acquiescence	1	25.0%	3	1	1	12.5%	3	30.0%
Adaptation	1	25.0%	4	2	2	25.0%	4	40.0%
Denial	1	25.0%	1	0	0	0.00%	1	10.0%

School Vignettes

This section will draw on survey and interview data to create a short description of the unique context related to teacher evaluation at each of the four schools. Detailed demographic information for each school is not included here but can be found in the school profiles in Chapter 4.

Charles

Charles was the smallest of the three traditional high school in the study (789 students) and had the highest participation in free or reduced price lunch (FRPL), at 46%, of the three traditional high schools. Of the four high schools in this study, teachers at Charles described a focus on testing results that was much more intense than the other three high schools. An intense focus on testing achievement may be the reason why Charles had the highest State Evaluation Effectiveness score at 9.3 above the district average. Conversely, Charles had the lowest Local Evaluation Effectiveness score, and the only negative score in the study, at -2.5. There were four focal teachers interviewed from Charles, two whom taught Math and two whom taught English.

Mrs. Ranier, an English teacher, stated that due to the testing focus, she felt the faculty at Charles was “analytical and cynical” and expressed surprise that even teachers in subject areas that have traditionally avoided state testing, such as the Art department, seemed this way. She stated that there was a disconnect between what the principal thought needed to be done to achieve good test scores and what teachers thought needed to occur, an issue Mrs. Ranier attributed to her principal’s lack of practical experience. The principal, Mrs. Warner, only spent three years in the classroom as a special education teacher before moving to administration. Mrs. Ranier felt she had to engage in a lot of required “cover your behind” activities that were expected of the teachers at her school, such as parent meetings and various types of documentation. She stated, “I really liked it better 10 years ago when I was trusted to do my job and do it well. And I think I did a better job because I was less anxious, there was less stress.”

Mr. Eagle, an English teacher who had taught previously out-of-state, but was in his first year at Charles, described a general feeling of being watched, both in the classroom and outside in meetings and other functions. While Mr. Eagle stated that the administration seemed very enthusiastic about his performance as a teacher, he acknowledged that there seemed to be an “invisible list of bad teachers” who were watched more frequently. He also noted an intense focus on student growth on tests. He had good scores his first semester with his seniors, which he attributed to pure luck as the majority of his semester had been spent helping students work on their senior projects rather than on topics and skills covered by the North Carolina Final Exam (NCFE), and he had already been approached about teaching an End of Course (EOC) class the following year due to this success. The suggested course reassignment signaled that the principal was willing to engage in the gaming strategy of moving teachers with the best records of producing growth to areas where testing stakes are higher (Cohen-Vogel, 2011; Grissom et al., 2012). The principal’s decision to pursue this highlights the testing focus of the school.

The views of the two focal English teachers were echoed by the two Math teachers who were interviewed for this study. Both Math teachers talked about the testing results-driven atmosphere of the school; however, the Math teachers did not seem to feel the same pressures and oversight that the English teachers expressed. There are three possible explanations for this. First, the closeness of the departments differed in professional relations, personal relations, and physical location. Mr. Robbins, the more veteran of the two Math teachers, described how the department consisted of established teachers who had taught for several years. Mr. Silver, the other focal Math teacher was the newest in the department and he had been there for several years and had student taught at Charles prior to being hired. The Math department was also very tight-knit and would meet on the weekends to play board games at each other's houses. In contrast, the English department consisted of many teachers who were new to the school, including both focal English teachers. The English department was also spread out across the school instead of being located in adjacent classrooms in one hallway like the Math department.

Secondly, the Math teachers were not observed by the principal in the study year, but were instead observed by an assistant principal who had a Math background. However, at Charles the administration observed different departments each year so subject area alignment was not guaranteed for Math teachers. In contrast, English teachers at Charles had never been observed by an administrator with English experience.

Finally, the Math teachers also received a lot of outside support in the form of a curriculum coach, while the English department declined assistance from their curriculum coach. This may mean that Math teachers relied more on feedback from the coach and their tight-knit group of peers whereas English teachers were primarily receiving feedback from the administrator. The stability of the Math department and the coaching support it received may help explain the different perceptions shared by teachers across the two subject areas.

There were no statistically significant differences in teacher perceptions of evaluation when comparing reflections on the prior year to the current year. Despite all four focal teachers acknowledging that there was a strong focus on raising test scores in their school, teachers did not demonstrate large rises on any of the testing-related questions. However, there were two themes where Charles demonstrated low means as compared to two of the other schools: Riley and Phoenix. Compared to teachers at these two schools, teachers at Charles seemed to feel more secure in their jobs despite evaluation policy and demonstrated little concern that evaluation would affect employment (prior year: M= 1.60, SD= 0.52; current year: M= 1.70, SD= 0.48) or result in the label of "bad teacher" (prior year: M= 1.80, SD= 0.92; current year: M= 1.70, SD= 0.48).

Central

Central was a mid-sized traditional high school (1,103 students) and was the least ethnically diverse with a student population that was 86% white. Central was above the district average for Local Effectiveness Scores and only slightly below for State Effectiveness Scores; however, the school had very low Local Condition Scores (-9.6) and even lower State Condition Scores (-22.5). These low Condition Scores indicate that teachers held negative views about evaluation policy at Central. These scores may be related to a unique situation where the Math and English departments seemed to have divergent experiences in regard to observation. The divergence was a result of the Math teachers being observed for several consecutive years by the main principal, Mr. Nichols, who had several years of experience as a Math teacher prior to becoming an administrator. In contrast, the English department was observed for several consecutive years by an assistant principal, Mr. Reward, who allowed teachers to complete their own evaluations because he struggled to operate a computer. There were three English teachers and two Math teachers who completed focal interviews from Central.

The three focal English teachers felt that their department was strong and did not require a lot of oversight. One teacher, Mrs. Williams, attributed Mr. Reward's assignment to the department as a testament to teacher skill and explained that she felt the main principal did such a good job at hiring that the teachers at Central did not need to be watched or evaluated. All three English teachers described evaluation under Mr. Reward in a similar way: the assistant principal would sit in for part of a class and then largely leave up the assigning of scores and comments to the observed teacher. Mrs. Williams described taking over the computer from Mr. Reward and typing up the evaluation for him; “And I think I’m fair,” she added.

One English teacher, Mrs. Hoard, was only required to have one evaluation in the study year and described how that observation did not occur with students. Instead, Mr. Reward observed her conducting a department meeting and filled out the observation rubric based on that observation. As a result, Mrs. Hoard was not formally observed teaching at all during the study year and a lot of areas were marked “not observed” on her observation form. She also noted that sections of the evaluation that were focused on students were rated by Mr. Reward despite an absence of students during the observation. Mrs. Hoard did not like that observations were conducted so haphazardly, but felt secure that the results of her evaluation would have no effect on her at all. Overall, the English teachers were very open about the experience and stated that the assistant principal was very good at many things, such as handling discipline and bus schedules, but that he was ineffective in many aspects of his position as an administrator. One of the Math teachers indicated in his interview that he had been observed once under Mr. Reward and described similar experiences, so the deficit, as described by the teachers, may have transcended subject areas.

Conversely, the Math department described very positive experiences with observation due to quality feedback from Mr. Nichols, the principal, who had a background as a Math teacher. The focal Math teachers described how Mr. Nichols would identify things in the observed lesson that may have otherwise gone unnoticed by the teacher or make suggestions that were practical and could be used to improve instruction. Both Math teachers expressed gratitude that they had an administrator who “knows the Math” and could effectively identify if something went wrong. Overall, the Math teachers seemed to feel that the feedback received from observations was useful and valid due to their administrator's background as a Math teacher.

Yet, there was a disconnect demonstrated between Central's high State Effectiveness scores and low State Condition scores. I asked the focal teachers to help explain how teachers at Central could have such high Effectiveness Scores yet have such a low impression of evaluation conditions at their school. Mrs. Proffitt, a Math teacher described how one of the assistant principals had been returned to classroom teaching at a different school after "making a mess of testing" the previous year with numerous, serious scheduling and protocol errors. Mrs. Proffitt suggested that this may have led to very low perceptions of testing in the previous year when the data for the Condition Scores were collected.

However, it may also be that the observation style of Mr. Reward was related to the Local Condition Scores. While only two subject areas were included in this study, Mr. Reward conducted at least one third of the observations at Central and presumably he had the same challenges with using technology and filling out the evaluation form for other subject areas as was reported for English and, in the past, at least one Math teacher. The interviews support this hypothesis as teachers were ranked highly at the school, often after having completed their own evaluation ratings, but were frustrated at the purported ineffectiveness of their evaluator. Additionally, teachers were very open in discussing Mr. Reward's tactics, so it may be that teachers who were not evaluated by him were also frustrated that some of their colleagues did not get evaluated in the same manner.

All of the teachers interviewed also stated that the overall consensus in their respective departments was that testing was a "fact of life" and that they wanted students to do well on tests; however, teachers overwhelmingly did not see value in the tests as sources of feedback or as valid measures of student gains. Mr. Augustus surmised, “I do not think we feel it is very useful or effective and it is a waste of time and a waste of resources and a waste of money.” He went on to describe how across the state, teachers of Math II courses had to administer two different EOCs that year and to uphold students’ impression that both tests counted toward grades, “[T]he giving of two tests, but only one of them we are going to have access to the data. That is, to me, not very useful and not very effective.” The unique situation Math II teachers were placed in is described in greater depth in the next chapter which examines differences based on individual characteristics such as subject area.

Similar to Charles, teachers at Central seemed to feel quite secure in their jobs and demonstrated little concern that evaluation would affect employment (prior year: M= 1.42, SD= 0.52; current year: M= 1.58, SD= 0.52) or result in the label of “bad teacher” (prior year: M= 1.67, SD= 0.65; current year: M= 1.83, SD= 0.84). There was also a significant difference when comparing responses from the prior to the current year on teachers’ use of feedback to modify practice, $t(11) = -2.46$, $p = 0.03$. However, this change only represented a shift from teachers on average agreement from “disagree” to “neither disagree or agree.” So, while teachers may have overall felt more inclined to use feedback in the study year, this change does not necessarily mean that teachers rely heavily on evaluation feedback for classroom planning.

Riley

Riley was the largest high school in the study (1,591 students) and was the most ethnically diverse. Riley also had the lowest percentage of students in FRPL (38%) and ranked very near the district average on all of the Condition and Effectiveness Scores. However, the aspect of Riley that is the most distinguishing is that it had a new principal, Ms. Jefferson, during the study year. According to the teachers, her approach to observation was very different from her predecessor's. Ms. Jefferson initiated significant changes in teacher classroom practice by requiring the submission of daily lesson plans on which she reviewed and commented. In addition, she immediately completed a full observation, complete with conferences and discussions about lesson plan submissions, of every teacher during the first month of the school year. For instance, Mr. Donaldson, an English teacher, was formally observed three times before our first interview in mid-October. He was not in a renewal cycle and was only legally required to have two snapshot observations (abbreviated observations covering only a few standards) in the study year. Given the size of the high school with around 100 teachers, observing each was an impressive accomplishment for Ms. Jefferson and required a considerable investment of her time.

Teachers seemed apprehensive about engaging in a study on evaluation given this new focus on observation, as evidenced by my conversations with the two focal teachers and from comments typed into the open response section of the survey. Nearly all Math teachers took the survey for this study (90.91% response rate), but none were willing participate in the interview. I spoke to the department head regarding the interview and he stated that the department had come to a decision not to participate in the interview because the atmosphere surrounding observation in the school had become "tense." In the English department, the survey participation rate was low (45.45% response rate) and the only teachers who volunteered to be interviewed were very secure veteran teachers who reported always receiving good marks on their respective evaluations.

The two focal English teachers were critical of the new administrator's approach to observation despite receiving glowing evaluations from Ms. Johnson. Their criticism centered on the fact that the new administrator equated the completion of paperwork with good teaching. Mr.

Donaldson stated that he observed “pushback from some quarters” of the teachers regarding the increased frequency and intensity of observations and that some teachers felt that there was an “entrapment factor” behind the approach.

Additionally, both focal teachers described how they had taken advantage of Ms. Jefferson’s tendency to focus on paperwork completion over content. For instance, Mrs. Macdonald had begun to modify how she submitted the lesson plans so that the process became more applicable to her own classroom practice. She noted that the principal had not challenged her alterations. Mrs. Macdonald also noted that after the first few lesson plan submissions she stopped receiving feedback and has since invested less time in the process. Meanwhile, Mr. Donaldson was not observed by Ms. Jefferson again between the first and second interviews and had started recycling lesson plans on his daily submissions, which he stated had gone unnoticed. However, Mrs. Macdonald said she was aware of several teachers who were being reprimanded for not completing paperwork, which seemed to support Mr. Donaldson’s “entrapment” theory.

Mrs. Macdonald was also skeptical that the very high marks she received on her initial observations were an accurate reflection of her teaching. She stated that the feedback she had received lacked substance and suggested that Ms. Jefferson was measuring her “enthusiasm and charisma” as a teacher rather than her ability. Additionally, Mrs. Macdonald felt that the fact that she met all deadlines on her paperwork submissions was instrumental in her doing well as the new principal seemed to value this over actual teaching ability.

The increased and intense focus on evaluation, tied to the submission of daily lesson plans, seemed to result in increased tension around observations at Riley. There were two statements on the complementary question set which had significant differences when comparing the prior to the current year. First, teachers were reportedly more likely to choose teaching strategies based on evaluation, *t*(14) = -1.87, *p* = 0.08. Secondly, teachers were reportedly less likely to direct focus on certain students based on evaluation, *t*(14) = 1.87, *p* = 0.08. However, the averages for both categories remained between the "disagree" and "neither agree nor disagree" ranges. So, these changes do not represent a reliance on evaluation for choosing strategies nor a total abandonment of using evaluation to direct focus on certain students. However, an increased awareness of the relationship between evaluation and the selection of teaching strategies might have resulted from the daily lesson plan requirement or from teachers' decisions to use certain strategies to meet observation requirements. Compared to other schools, the mean responses on the survey were not exceptionally high or low. For the most part, responses from Riley teachers fell in the middle of the means for all of the question themes on the survey.

There was no change at all in teachers' intentions to use testing data to modify instruction as compared between the prior year and the current year as measured in the survey. The focal teachers expressed that they felt that testing would not factor into their evaluation in a way that was different than before Ms. Jefferson came to the school. However, this survey was administered at the beginning of the school year before a testing cycle had been completed. Both focal teachers commented that the English department was strong, and they were surprised that Riley was slightly below the district average for standard six of the evaluation. Mrs. Macdonald explained that the school was located in a higher socio-economic area than the other schools and stated that students were "high testers" at the school. The school scores were always near the top of the district, so this slightly than lower average may have resulted from the way growth was calculated rather than from low student scores. Mrs. Macdonald also stated that despite Riley usually scoring at the top for the district, she still paid attention to testing and remembered a time when the Biology teachers had low scores which resulted in reprimands and teachers being sent away for additional training, which was something she thought about every testing season when she began to review for exams.

Phoenix

Phoenix has a unique context as an alternative school serving students who did not experience success in traditional high schools. As an alternative school it was the smallest in the study (134 students) and had the highest rate of FRPL (87%) when compared to the traditional high schools. The Local Condition Score and Local Effectiveness Score were both slightly above the average for the district. However, the State Condition Score was low (-16) and the State Effectiveness Score was the lowest in the study (-13).

When asked about attitudes towards testing at his school, Mr. Brown, a second-year English teacher who was in his first year at Phoenix, surmised, "[There] are bigger fish to fry here [than academics], especially on the social and emotional level. Some of these kids are dealing with a lot [which] matters more in the grand scheme of things. Some of these kids need to have social skills as opposed to knowing how to take a test." During our interviews, Mr. Brown often reflected on the difference in his experiences between his first year at his first school, which was located in Tennessee, and his second and current year at Phoenix. He felt the difference was related to the school administration. At his previous school, Mr. Brown felt anxiety about his evaluations and perceived extreme pressure on teachers to achieve high test scores. He stated that everything in his observations seemed to be linked back to state testing and described his first year as being in a classroom that was micromanaged by policies meant to increase student achievement on tests.

I observed a noticeable difference in Mr. Brown's anxiety about testing between his first and second interview. In the first interview, Mr. Brown expressed nervousness about how testing would play out in the unique new classroom context he now taught in as he previously had faced retribution if his students did not perform well. By the second interview and second semester, Mr. Brown seemed to have accepted that testing did not matter in the same way at Phoenix as it did at his previous school. Mr. Brown stated that the administration at Phoenix did not have low expectations, but instead, "They understand, in fact, they really understand what the teachers are dealing with, like how a classroom looks. And I feel like their perceptions, they align maybe pretty well with the teachers. Maybe that is why teachers feel that the scores are pretty accurate. I felt like mine were pretty accurate."

Mrs. Street is a veteran English teacher who previously served as a curriculum coach at Phoenix. She returned to the classroom to finish out her teaching career with a few years left before retirement. She talked extensively about the autonomy that was afforded to teachers at Phoenix, which was something she felt did not occur at other schools. Mrs. Street stated that the principal of Phoenix allowed and encouraged teachers to try new things to reach the unique population of students they served. Mrs. Street described how she had felt in other schools, particularly during observations, "I would be very nervous to try new or out of the ordinary [methods]. I would stick to something more scripted, something tried and true. Here we have the freedom. We are not going to be marked down for trying a strategy or trying something with students and it fails."

The focal teachers interviewed for the study also brought up the culture of Phoenix, which Mrs. Street terms as being one of "learning and growing" where every teacher is willing to accept feedback from the others. Mrs. Street explains, "[F]eedback is necessary… it is not a bad thing or just a good thing, it is how can we all learn from each other." Mr. Forest described conversations he had with colleagues from other schools about school climate and wondered if the better relationship between teachers and administrators he perceived at his school was the result of accessibility. "Since we are such a small staff here, I feel like I can walk into [my administrator's] office at any time and I wonder if my colleagues at other schools feel the same way." Mr. Forest's statement does seem to contrast statements from teachers at the other focal schools who stated that administration did not enter classrooms aside from evaluations or meet informally.

The mean for Phoenix was fairly higher on the statement "choosing teaching strategies based on what evaluated on" when compared to the other schools in the study. Additionally, the mean for this statement rose when comparing the prior to the current year, though the change was not statistically significant. Perhaps this increased focus on strategies is related to the unique student population of Phoenix and a school climate where experimentation is encouraged. Additionally, there was a significant, positive difference for the statement "use test data to modify classroom practice" when comparing the prior to the current year with the mean rising, $t(4) = -2.24$, $p = 0.09$. Nothing in the interviews indicated that policy changes at the school drove these changes in means. However, the small sample size may indicate that this change simply reflects the personal resolve of those teachers who completed the survey.

Overall, the focal teachers' descriptions of the context and climate of Phoenix seemed to match the conditions and effectiveness scores used to select the school for the study, which boasted high Local Condition scores and Local Effectiveness Scores slightly above the district averages. It is interesting to note that three of the themes in which large rises were demonstrated at Phoenix were among the same as those at Riley: "modifying practice using feedback from evaluation," "choosing teaching strategies based on evaluation," and "using observation data to modify classroom practice." A very large rise was also seen in the theme "using testing data to modify classroom practice." Yet, Phoenix teachers had very low perceptions of State Conditions and State Effectiveness with scores of -16 and -13 below the district average, respectively, which would lead one to believe that teachers did not rely on testing feedback to modify instruction and that even with the growth model, the school underperformed compared to the majority of others in the district. As referenced before, it is unclear if these results were driven by the addition of two new-to-Phoenix teachers to the current year question set as opposed to the prior year or if these statements merely reflected a continuing dedication to "learning and growing" that was espoused by Mrs. Street and supported by statements from the other two focal teachers.

Do Evaluation Conditions and Effectiveness Matter?

Overall, the measures used to select schools may not have been effective in identifying differences of context because no significant differences in survey results were found among the four schools. Additionally, the effects may be understated due to the small sample size. However, contextual differences may have impacted the relationship between evaluation and teacher practice in schools. The interviews suggested that such a relationship may be more related to the type of individual who is conducting evaluations and how evaluation is related to the climate and policy focus of the school.

For instance, it may be important to consider whether evaluation is even stressed by school administration and if so, which parts of the evaluation are emphasized? Other studies have demonstrated that teachers respond to evaluation through the lens of their administration and so the way in which principals choose to focus on evaluation may influence how teachers perceive the policy (Reinhorn, Johnson, & Simon, 2017). Do administrators stress the observation or the testing component more? Are there certain aspects of each component that receive more focus than others? Additionally, in regard to the observation component of evaluation, it may be important to consider the subject area background, the skill of the observer, and the foci or values administrators bring in approaching evaluation. These are specific considerations that were not taken into account in the way in which the Evaluation Condition Scores were calculated. For instance, the two teachers interviewed at Riley questioned the proficiency of the new principal who stressed observation as important, but seemed to connect this evaluation component with paperwork and check boxes rather than providing meaningful feedback. Observer proficiency certainly was reported as an issue for the English teachers at Central who described cases of writing their own observations or not being observed teaching at all. While subject area will be discussed in more depth in the next chapter, the Math teachers who were observed by administrators with Math backgrounds expressed gratitude that their observer had proficiency in the subject area observed. Additionally, teachers, such as Math teachers at Charles, who receive quality feedback from outside sources such as coaches, may value coaches' feedback over formal evaluation feedback.

The interviews also revealed some considerations not captured by the Evaluation Effectiveness Scores. First, the local scores which were based on observation were clustered closely around the district mean with a range of -1.45 to 1.09. This trend was consistent in all the schools in the district sample; therefore, a wide range of scores in this category was unavailable. For the state-based score, which was based on standard six, also known as the student growth standard, the range was much wider, -14.41 to 8.81. Charles and Central were above the district mean while Riley was slightly below, and Phoenix had the worst performance among high schools in Broadville. However, the teachers at Phoenix spoke at length about how the school administration granted them autonomy from school-based policy related to testing and encouraged an atmosphere of experimentation to help the students at the school succeed in other, perhaps more meaningful ways. While Phoenix teachers ranked the lowest in Effectiveness scores for both local and state components, they had the highest Local Condition Score of the focal schools. The high score is not surprising given the level of autonomy granted to teachers there. Therefore, it is possible that the conditions under which teachers receive evaluation scores have more of an association with teacher perceptions of evaluation than the actual evaluation scores that are received by those teachers.

Evaluation Scenarios

Overall, the vignettes reveal three types of scenarios related to evaluation. The first is the technocratic scenario. Giroux (1985) argued that a technocratic approach to policy reduces teacher autonomy by attempts to regulate and control behavior. For instance, Mr. Brown, the new English teacher at Phoenix, described these conditions in his previous school while contrasting his two teaching experiences. This technocratic approach is also evidenced to a large extent in Ms. Jefferson's approach at Riley. According to teachers, Ms. Jefferson focused more on observation than the testing component with a formalities-driven approach to observation based on controlling teacher behavior through lesson plan submission and oversight with possible reprimand through observation. An adherence to procedure was valued rather than quality of work. Additionally, technocratic approaches regarding the testing component of evaluation may have already been in place prior to Ms. Jefferson's hire as evidenced by Mrs. Macdonald's fear of what happened to the Biology teachers following poor testing performance. Firestone (2014) argued that duress is the opposite of autonomy; it appears that teachers at Riley may have been experiencing some duress around evaluation. Therefore, evaluation at Riley, according to teachers, is fully under a technocratic model.

A technocratic approach is also evident, but to a lesser extent, at Charles, particularly in the English department. Scholarship has demonstrated that conditions for motivation in schools include realistic workloads, administrative support, and operating in systems that are not overly punitive (Firestone & Pennell, 1993; Firestone & Rosenblum, 1988). Mrs. Ranier described a challenging workload and lack of support from administration that expected teachers to cover their own "behinds" and specifically referenced how such a school climate impacted the motivation of teachers. All of the teachers at Charles stressed the results-driven nature of the school in regard to testing, a condition Giroux (1985) describes as being technocratic. The Math teachers may have appeared to be more aligned with a test-driven approach due to having group buy-in towards this policy. Additionally, the success of Math teachers was also aided by the creation of Introductory Math courses which were required of all students and meant to increase student success on the Math I test. Because the course helped ease the burden of teaching Math concepts by spreading the curriculum across two semesters instead of one, the Introductory Math course may have helped create this policy buy-in for Math teachers. So, Charles is also a school that operates under a technocratic model, particularly in regard to the testing component of evaluation.

The second evaluation scenario is the Autonomous and Self-Efficacious Scenario. In this scenario, teachers are able to operate under a system of internal rewards where improvement is driven by assessment, feedback, training, and professional development while evaluation contributes to rewarding conditions (Firestone, 2014). A clear example of this can be seen at Phoenix, where teachers felt supported by administration and worked in an atmosphere of learning and improvement guided by their administrator. While teachers did not necessarily view evaluation, and particularly the testing component, in a positive manner, they were satisfied overall with the way observations were conducted at their school and with the administration's approach to the policy. However, this did not mean that the administration at Phoenix was hands-off. Aside from Riley, Phoenix was the only school that required submission of lesson plans. It was also the only school where teachers mentioned administrator presence in the classroom aside from evaluation. Overall, while the staff at Phoenix did adhere to the policy requirements of teacher evaluation, the policy did not define instruction and instead teachers and administrators were free to work together to create their own definitions of success in more meaningful and supported ways. Therefore, Phoenix serves as an excellent example of teachers working under conditions of both autonomy and self-efficacy as allowed in a non-traditional high school.

To a lesser extent, the Math departments at Charles and Central exhibited some tendencies consistent with this scenario. The Math department at Charles exhibited a level of autonomy not present in the English department, perhaps because of the overall closeness and stability of the department and perhaps because there was greater buy-in from the Math teachers regarding the value of a testing-focused curriculum and technocratic policies compared to the English department. For these reasons, Charles more appropriately fits into the Technocratic Scenario as previously described. Likewise, Math teachers at Central experienced some tendencies consistent this scenario, partially driven by having an administrator with a Math background resulting in a mutual recognition of competence in the subject. However, Central more readily fits into the final scenario described below.

The final scenario is Consensus Lacking. Cohen (2011) describes how a lack of consensus in an educational context can increase uncertainty and dispute. The teachers at Central certainly described themselves as being autonomous and self-efficacious and teachers in both departments were dismissive of both the observation and testing components of the evaluation as evidenced by the Evaluation Condition Scores and the interview data. The evaluation scores received by teachers at Central indicated that teachers were accomplishing a "good job" by those measures, yet when it came to evaluation there was a lack of consensus regarding the policy. Math teachers seemed to find some validity in their own personal observations, which were conducted by a former Math teacher, but did not seem to find value in the observation policy or process as a whole. Meanwhile, the English teachers' evaluations were conducted in a manner more consistent with the Wild West, where teachers were sometimes not actually observed teaching or essentially observed themselves. However, neither department found value in the testing component of evaluation. Overall, there was a lack of consensus and an attitude of even disdain toward evaluation at Central that was supported by the way in which administration approached the policy.

Additionally, the interview data presented earlier in this chapter in Table 13 offer some support for the scenarios presented above, though the results should be interpreted with caution due to the small size of the samples at each school. For instance, Phoenix had the highest occurrence of the reform typology of acquiescence (which occurred when a teacher made a statement that indicated an acceptance of the policy, which could be with reluctance but without protest) of the four focal schools with 50% of the interviews containing statements that indicated acquiescence. The second highest occurrence was found at Central where 30% of interviews included statements indicating acquiescence. It could be that teachers at Phoenix were more likely to demonstrate acquiescence to evaluation policy because it was unlikely to interfere with their classroom lives under the Autonomous and Self-Efficacious scenario.

All of the teachers who demonstrated acquiescence at Central were English teachers, whose department often completed observations with little input from the administrator. Again, the way evaluations were conducted at Central may have led to less interference in the classroom lives of teachers, but did little to improve teaching conditions at the school. Phoenix teachers also mentioned internal motivation more frequently by percentage of interview, at a rate of 66.7%. Conversely, at Riley, the school under the Technocratic scenario, there were internal motivation statements in only 25% of the interviews. Again, this is sensible given that individuals who are allowed to act autonomously and practice self-efficacy tend to be internally motivated, whereas teachers in Riley were operating under a system of external threats and rewards.

In summary, the survey results did not indicate that differences in evaluation conditions or effectiveness as measured in this study affected teacher perceptions of the relationship between evaluation and practice. However, teachers' statements during the interview phase suggest that conditions, particularly related to administrator implementation of the policy and expectations around effectiveness scores, do matter. Specifically, more work is needed to parse out how components of evaluation conditions impact teachers differently.

CHAPTER 7: Individual-Level Characteristics and Teacher Evaluation

This chapter addresses the question of whether individual-level teacher factors are associated with differences in teacher motivation, experiences with feedback, and work decisions related to teacher evaluation. Differences are examined by comparing teacher reflections on the prior and current year complementary survey questions as well as by examining differences between groups. Specifically, three individual-level differences are examined: reported licensure status (provisional vs. professional), years of experience (seven years or fewer vs. eight years or more), and subject area (Math vs. English). In this section, I present an analysis of survey and interview data to examine the relationships between each of the three individual-level factors and teacher perceptions of evaluation. Throughout, I explain how I found individual-level characteristics to be related to the perceptions of teachers at the four focal schools. Additionally, I present some alternative explanations for the differences that emerge.

Reported Licensure Level

Survey

Two licensure types were reported among survey participants: provisional and professional. It is important to note that all teachers who were provisionally licensed were subjected to full observation cycles each year which consisted of three full-length observations and a peer observation assessed by all five observation standards along with standard six, which is the student growth standard. All observations are supposed to include conferencing between the observer and the observed teacher. Teachers with a professional license may have either a full cycle or an abbreviated cycle (two snapshot observations evaluating three observation standards plus the growth standard), dependent on when the teacher began teaching in North Carolina and whether or not their license is up for renewal in a given year.

Quantitative analysis revealed some differences between licensure levels for both the prior year and current year question sets. An independent samples t-test was conducted to determine the relationship between teacher licensure level and perceptions of the teacher evaluation process in the prior year. An analysis of teachers' reflections on the previous year yielded two significant results when comparing teachers who reported provisional licensure to those who reported professional licensure. First, provisionally licensed teachers were more likely to have stated that they chose curriculum in anticipation of evaluation than professionally licensed teachers, $t(40) = 1.93$, $p = 0.06$. Additionally, provisional teachers were also more likely to agree that they directed focus on certain students based on evaluation compared to professionally licensed teachers, $t(40) = 1.96$, $p = 0.06$). Unfortunately, these statements were about evaluation at large so it is unclear whether these responses may have been associated differentially if examined separately by observation or testing (see Table 15).

An independent samples t-test was also used to examine the relationship between teacher licensure level and perceptions of the teacher evaluation process in the current year. Three significant differences emerged between provisionally and professionally licensed teachers. Overall, provisional teachers were more likely to state that they anticipated modifying practice in anticipation of an evaluation as opposed to professionally licensed teachers, $t(42) =1.96$, $p = 0.06$. Provisionally licensed teachers were also more likely to be concerned that an evaluation would label them a bad teacher as opposed to professionally licensed teachers, $t(42) = 1.84$, $p = 0.07$ and, as in the prior year question set, provisionally licensed teachers were more likely to direct focus on students based on what they will be evaluated on as opposed to professionally licensed teachers, $t(42) = 3.02$, $p < 0.01$.

Table 15

Independent Sample T-Test of Survey Themes by Reported Licensure Level

		Provisional			Professional					
Survey Themes		M	SD	n	M	SD	n	95% CI	t	df
Modify practice in anticipation of an evaluation	Prior	3.00	1.15	10	2.50	1.24	32	-0.40, 1.40	1.13	40
	Current	3.36	1.21	11	2.52	1.25	33	-0.02, 1.72	1.96*	42
Modify practice using feedback from evaluation	Prior	3.20	1.40	10	2.69	1.15	32	-0.37, 1.40	1.17	40
	Current	3.36	1.29	11	3.15	1.03	33	-0.56, 0.99	0.55	42
Have concern evaluation affects employment	Prior	2.30	1.25	10	1.66	1.00	32	-0.15, 1.44	1.63	40
	Current	2.36	1.12	11	1.88	1.08	33	-0.28, 1.25	1.28	42
Have concern evaluation labels as a bad teacher	Prior	2.30	1.25	10	1.72	0.96	32	-0.17,1.34	1.56	40
	Current	2.45	1.04	11	1.79	0.86	33	-0.06, 1.21	1.84*	42
Have concern evaluation does not reflect competency	Prior	2.50	1.27	10	2.56	1.22	32	-0.96, 0.84	-0.14	40
	Current	2.45	1.04	11	2.45	1.23	33	-0.76, 0.76	0.00	42
Choose curriculum based on what evaluated on	Prior	3.20	1.23	10	2.31	1.28	32	-0.04, 1.82	1.93*	40
	Current	2.80	1.14	10	2.45	1.23	33	-0.54, 1.23	0.79	41
Choose teaching strategies based on what evaluated on	Prior	3.10	1.20	10	2.59	1.27	32	-0.41, 1.42	1.12	40
	Current	3.55	1.21	11	2.85	1.23	33	-0.16, 1.56	1.64	42
Direct focus on certain students based on what evaluated on	Prior	3.10	1.20	10	2.28	1.14	32	-0.03, 1.67	1.96*	40
	Current	3.27	1.10	11	2.18	1.01	33	0.36, 1.82	3.02***	42
Use test data to modify classroom Practice	Prior	3.40	1.08	10	3.19	1.18	32	-0.63, 1.06	0.51	40
	Current	3.36	0.92	11	3.33	1.14	33	-0.74, 0.80	0.08	42
Use observation data to modify classroom practice	Prior	3.00	1.25	10	2.75	1.16	32	-0.62, 1.12	0.58	40
	Current	3.00	1.34	11	3.24	1.15	33	-1.08, 0.60	-0.58	42
Feel evaluated fairly	Prior	3.70	0.82	10	3.88	1.13	32	-0.96, 0.61	-0.45	40
	Current	4.10	0.57	10	4.15	0.57	33	-0.46, 0.36	-0.25	41
Feel last year will impact current Year	Prior	3.00	1.14	10	2.34	1.13	32	-0.22, 1.53	1.52	40
	Current									

Note. Scale for Survey: Strongly Disagree 2- Disagree 3- Neither Agree nor Disagree 4- Agree 5- Strongly Agree
* = p< 0.1, **= p< 0.05, *** = p < 0.01

Interview

The data in Table 16 present the interview case for each code by reported licensure status. The frequency is the number of interviews that included at least one occurrence of the code, which was used rather than frequency of interviewees due to the small "n" of these data when divided into categories. Each interviewee gave two interviews for the study so the "n" reported equals twice the number of interviewees at a school. The percent of occurrence for the interview in each category is included to allow for comparisons between the two groups.

Table 16

Occurrence of Codes in Interviews by Licensure Status

	Provisional (n= 6)		Professional (n= 22)	
Codes	Frequency	Percentage	Frequency	Percentage
Motivation				
Internal	4	66.7%	10	45.5%
External	2	33.3%	8	36.4%
Observation Feedback	4	66.7%	19	86.4%
Negative	4	66.7%	18	81.8%
Positive	2	41.7%	7	31.8%
Testing Feedback	4	66.7%	19	86.4%
Negative	4	66.7%	19	86.4%
Positive	1	16.7%	1	4.5%
Work Decisions				
Strategy/How Taught	1	16.7%	7	31.8%
Curriculum/What Taught	2	33.3%	4	18.2%
Who is Taught	0	0.0%	2	9.1%
Response to Reform				
Acquiescence	1	16.7%	7	31.8%
Adaptation	1	16.7%	7	31.8%
Denial	1	16.7%	2	9.1%

Discussion

The survey results indicated that provisional teachers may have been more likely than professionally licensed teachers to modify their practice in anticipation of an evaluation, specifically in regard to choosing curriculum and directing focus on students. Support is found in the focal interviews where a much larger percentage of interviews from provisionally licensed teachers referenced changes in curriculum based on evaluation as compared to professionally licensed teachers (33.3% versus 18.2%); however, the sample size of the interview data is too small to investigate with inferential statistics. Likewise, it is unsurprising that provisionally licensed teachers, who in general have less experience and perhaps less confidence in their instructional decisions, are more likely to fear being labelled negatively on an evaluation as opposed to more experienced, and possibly more confident, fully-licensed teachers. Provisional teachers, who do not have any sort of tenure protection or any prospect of receiving such under the current policies, also have more at stake if poor evaluation results are received. With this in mind, it is possible that evaluation may differentially motivate provisional teachers to change practice in an attempt to perform better on the evaluation measures.

There are other possible explanations for the differences found in the survey data. For instance, professionally licensed teachers, who in general have more work experience than provisionally licensed teachers (the exception being teachers who transfer from out of state and receive provisional licenses), may value something else aside from evaluations when it comes to classroom practices such as choosing curriculum or directing focus on certain students. There may be evidence for provisional teachers being more favorable to evaluation feedback than professional teachers in the focal interviews where there were fewer mentions of the negative aspects of both observation (66.7% versus 81.8%) and testing (66.7% versus 86.4%) from provisional teachers as opposed to professional. Provisional teachers may have a greater likelihood to focus on one or both components of evaluation to aide in these classroom practices because they have had less exposure to other types of guidance, and perhaps less confidence in identifying good practice, therefore seeing the values in evaluation as suitable guiding principles. However, it is difficult to draw support and conclusions from the focal interviews due to the sample only including three provisionally certified teachers, two of which were from Phoenix and one from Charles.

Seven-Year Status

Survey

As previously mentioned, changes in evaluation policy over recent years require all teachers with seven or fewer years of experience in North Carolina to be observed on a full evaluation cycle. Teachers who fall into the category of having seven or fewer years of experience may have either provisional or professional licenses; therefore, a different sample of teachers was included when the data was examined by the years of experience instead of by licensure level. So, independent sample t-tests were conducted using teacher seven-year status instead of reported licensure level (Table 17). No significant differences were found between teachers who had seven or fewer years of experience and teachers who had eight or more years of experience when analysis was run on the prior year question set; however, four significant differences between the two groups were found in the analysis of the current year question set.

First, the seven years or fewer teachers were more likely to both modify practice in anticipation of an evaluation compared to the eight years or over group, $t(42) = 2.28$, $p = 0.03$, and were more likely to modify practice using the feedback of an evaluation as opposed the eight years or over group, $t(42) = 1.81$, $p= 0.08$. Specifically, the seven years or fewer teachers were more likely to use test data to modify classroom practice than the eight years or over teachers, $t(42) = 2.49$, $p = 0.02$. Among those practices that teachers stated they would modify, seven years or fewer teachers were more likely to choose teaching strategies based on evaluation than the eight years or over teachers, $t(42) = 2.92$, $p = 0.01$ and were more likely to direct focus on certain students than eight years or over teachers, $t(42) = 3.26$, $p < 0.01$.

Table 17

Independent Sample T-Test of Survey Themes by Seven-Year Status

		7 or fewer			8 or more					
Survey Themes		M	SD	n	M	SD	n	95% CI	t	df
Modify practice in anticipation of an evaluation	Prior	3.13	0.99	8	2.50	1.26	34	-0.34, 1.59	1.31	40
	Current	3.56	1.01	9	2.51	1.27	35	0.12, 1.97	2.28**	42
Modify practice using feedback from evaluation	Prior	3.13	1.55	8	2.74	1.14	34	-0.58, 1.36	0.81	40
	Current	3.78	0.97	9	3.06	1.08	35	-0.08, 1.52	1.81*	42
Have concern evaluation affects employment	Prior	1.88	1.36	8	1.79	1.07	34	-0.81, 0.97	0.18	40
	Current	2.22	1.09	9	1.94	1.11	35	-0.56, 1.11	0.68	42
Have concern evaluation labels as a bad teacher	Prior	2.00	1.31	8	1.82	1.00	34	-0.67, 1.02	0.67	40
	Current	2.22	0.97	9	1.86	0.91	35	-0.33, 1.06	1.06	42
Have concern evaluation does not reflect competency	Prior	2.75	1.17	8	2.50	1.24	34	-0.72, 1.22	0.52	40
	Current	2.56	1.01	9	2.43	1.09	35	-0.69, 0.94	0.32	42
Choose curriculum based on what evaluated on	Prior	2.75	1.49	8	2.47	1.29	34	-0.77, 1.27	0.54	40
	Current	2.88	0.99	8	2.46	1.25	35	-0.54, 1.37	0.89	41
Choose teaching strategies based on what evaluated on	Prior	3.00	1.31	8	2.65	1.25	34	-0.65, 1.36	0.71	40
	Current	3.89	0.93	9	2.80	1.23	35	0.30, 1.88	2.92***	42
Direct focus on certain students based on what evaluated on	Prior	2.75	1.28	8	2.41	1.18	34	-0.62, 1.29	0.72	40
	Current	3.44	0.88	9	2.20	1.05	35	0.47, 2.02	3.26***	42
Use test data to modify classroom practice	Prior	3.50	1.07	8	3.18	1.17	34	-0.59, 1.24	0.58	40
	Current	3.89	0.60	9	3.20	1.13	35	0.12, 1.26	2.49**	42
Use observation data to modify classroom practice	Prior	3.13	1.36	8	2.74	1.14	34	-0.55, 1.33	0.55	40
	Current	3.67	1.12	9	3.06	1.19	35	-0.28, 1.50	1.39	42
Feel evaluated fairly	Prior	3.75	0.46	8	3.85	1.16	34	-0.95, 0.75	-0.24	40
	Current	4.13	0.35	8	4.14	0.60	35	-0.47, 0.43	-0.08	41
Feel last year will impact current year	Prior	2.88	1.55	8	2.41	1.13	34	-0.50, 1.43	0.97	40
	Current									

Note. Scale for Survey: Strongly Disagree 2- Disagree 3- Neither Agree nor Disagree 4- Agree 5- Strongly Agree
* = p< 0.1, **= p< 0.05, *** = p < 0.01

Interview

The data in Table 18 present the interview case for each code by seven-year status. The frequency is the number of interviews that included at least one occurrence of the code, which was used rather than frequency of interviewees due to the small "n" of these data. Each interviewee participated in two interviews for the study so the "n" reported equals twice the number of interviewees at a school. The percent of occurrence among interviews in each category is included to allow for comparisons across groups.

Table 18

Occurrence of Codes in Interviews by Seven-Year Status

	Taught Seven Years or Fewer (n= 12)		Taught Eight Years or More (n= 16)	
Codes	Frequency	Percentage	Frequency	Percentage
Motivation				
Internal	6	50.0%	7	43.8%
External	4	33.3%	6	37.5%
Observation Feedback	10	83.3%	10	62.5%
Negative	10	83.3%	10	62.5%
Positive	5	41.7%	4	25.0%
Testing Feedback	9	75.0%	11	68.8%
Negative	9	75.0%	11	68.8%
Positive	1	8.3%	1	6.3%
Work Decisions				
Strategy/How Taught	6	50.0%	2	12.5%
Curriculum/What Taught	3	25.0%	3	18.8%
Who is Taught	1	8.3%	1	6.3%
Response to Reform				
Acquiescence	2	16.7%	6	37.5%
Adaptation	4	33.3%	4	25.0%
Denial	1	8.3%	2	12.5%

Discussion

Other studies have demonstrated that novice teachers are less effective than more experienced teachers (Clotfelter, Ladd, & Vigdor, 2007; Rockoff, Jacob, Kane , & Staiger, 2011; Wayne & Youngs, 2003). Yet, newer teachers make rapid gains early in their careers (Boyd, Lankford, Loeb, 2008; Rockoff, 2004) and improve most rapidly in schools with higher socioeconomic status and higher schoolwide VAM scores (Loeb, Kalogrides, and Beteille, 2012). Three trends emerged from the combination of data from the survey and interview as separated by seven-year status which could offer explanation for the changes seen early in teacher careers. First, teachers who had seven years or fewer of experience were more likely to state they would change practice in anticipation of an evaluation than those with more experience. Similarly, a study by Sun, Mutcheson, & Kim (2016) demonstrated that early career teachers were more likely to use evaluation feedback to improve their practice, a finding that was reflected in the results of this dissertation.

Possible explanations include that teachers with fewer than seven years' experience are observed much more frequently and, due to lacking tenure, job retention is more closely tied to evaluation if a reduction in workforce were enacted. Therefore, it could be that more exposure to evaluation led to a greater awareness of changing practice to meet observation targets or that the higher stakes of evaluation led to such changes.

Second, teachers in the seven years or fewer category were more likely to state they would use feedback from an evaluation in general and specifically, feedback in the form of test data, than teachers with more experience. Again, this reliance could be due to increased observations and/or the greater stakes attached to lacking tenure. Interestingly, in the interview data, the frequency of occurrences of interviews mentioning feedback from observation in a positive manner was much higher among the seven years or fewer group versus the eight years or more (41.7% versus 25.0%), which suggests that there may be a relationship between the frequency of formal observations and the perceived value of the feedback received. Teachers who were interviewed often cited the limited number of observations conducted as influencing their ability to use observation feedback. So, it may be that those who are observed more frequently see greater value in the experience as a source of feedback.

Third, teachers in the seven years or fewer category were more likely to state that they changed their practice, specifically by choosing teaching strategies and directing focus on students based on evaluation than teachers with more experience. The interview data were supportive of this as interviews from teachers in the seven years or fewer category more frequently contained statements referencing a change in teaching strategies based on evaluation as opposed to the eight years or more group (50.0% versus 12.5%). Because teachers in the seven years or fewer group are evaluated on all six standards instead of four, it is also possible that teachers were simply trying to meet some of the standards with superficial changes. Superficial changes in practice dominated teachers' descriptions of changes in teaching strategies as discussed in Chapter 5. It is also possible that teachers who fell into the eight years or more category who received less observations and were more likely secured with tenure simply felt less external pressure from the evaluation policies. Again, interview data supports this as interviews from teachers in the seven years and fewer group were less likely to demonstrate a reform typology of acquiescence than those in the eight years and over group (16.7% versus 37.5%) indicating that there was less acceptance of the policy among those teachers subjected to full evaluation cycles.

Another possible explanation for the differences between the seven years or fewer and eight years or more groups may be unrelated to the pressure or frequency of evaluation, but due instead to the growth model that has been employed by administrators using the observation rubric, which was a common complaint among teachers who questioned the validity of the observation instrument as discussed in Chapter 5. It is possible that teachers with fewer years of experience are simply rated lower than those who have been evaluated across a longer time span because administrators may feel policy pressure to score newer teachers low initially in order to demonstrate growth later on. However, it should also be noted that the focal interview teachers in the seven years or fewer category were spread evenly across three of the focal schools: Charles, Phoenix, and Central, and there were no teachers from Riley represented in that sample. Such unbalance may help explain the differences seen between groups in the interview data.

Subject Area

Survey

The third individual-level characteristic of interest was subject area: whether the teachers taught Math or English. Independent samples t-tests were run on both sets of questions from the complementary question set, which asked teachers to reflect on statements regarding the prior and then the current year, to identify differences between Math and English teachers.

Differences between subject areas emerged in four themes between both question sets (Table 19). For both questions sets, English teachers were more likely to state they modified practice in anticipation of evaluation when compared to Math teachers (prior year: $t(40) = 2.07$, $p = 0.05$; current year: $t(42) = 3.18$, $p < 0.01$). While English teachers were more likely than Math teachers to report changing practice in anticipation of an evaluation, Math teachers appear significantly more likely to say that they would use observation feedback to modify classroom practice when compared to English teachers, *t*(40) = -2.09, *p*= 0.04. Similarly, Math teachers were significantly more likely to say the prior year's evaluation would impact current year classroom practice when compared to English teachers, *t*(40) = -1.18, *p* = 0.07.

Table 19

Independent Sample T-Test of Survey Themes by Subject Area

		English			Math					
Survey Themes		M	SD	n	M	SD	n	95% CI	t	df
Modify practice in anticipation of an evaluation	Prior	3.06	1.26	18	2.29	1.12	24	0.02, 1.51	2.07**	40
	Current	3.37	1.12	19	2.24	1.20	25	0.41, 1.84	3.18**	42
Modify practice using feedback from evaluation	Prior	2.50	1.34	18	3.04	1.08	24	-1.3, 0.21	-1.45	40
	Current	3.16	1.21	19	3.24	1.01	25	-0.76, 0.60	-0.25	42
Have concern evaluation affects employment	Prior	1.78	1.11	18	1.83	1.13	24	-0.76, 0.65	-0.16	40
	Current	2.00	1.05	19	2.00	1.16	25	-0.68, 0.68	0.00	42
Have concern evaluation labels as a bad teacher	Prior	2.00	1.24	18	1.75	0.90	24	-0.42, 0.92	0.76	40
	Current	1.89	0.81	19	1.96	1.02	25	-0.64, 0.51	-0.23	42
Have concern evaluation does not reflect competency	Prior	2.44	1.20	18	2.63	1.25	24	-0.95, 0.59	-0.47	40
	Current	2.42	1.07	19	2.48	1.09	25	-0.72, 0.60	-0.18	42
Choose curriculum based on what evaluated on	Prior	2.72	1.27	18	2.38	1.35	24	-0.48, 1.18	0.85	40
	Current	2.63	1.07	19	2.46	1.32	24	-0.58, 0.93	0.47	41
Choose teaching strategies based on what evaluated on	Prior	3.00	1.33	18	2.50	1.18	24	-0.29, 1.29	1.29	40
	Current	3.16	1.21	19	2.92	1.29	25	-0.53, 1.01	0.62	42
Direct focus on certain students based on what evaluated on	Prior	2.67	1.46	18	2.33	0.96	24	-0.48, 1.14	0.84	40
	Current	2.53	1.12	19	2.40	1.16	25	-0.58, 0.83	0.36	42
Use test data to modify classroom practice	Prior	3.17	1.10	18	3.29	1.20	24	-0.85, 0.60	-0.35	40
	Current	3.21	1.03	19	3.44	1.12	25	-0.90, 0.44	-0.70	42
Use observation data to modify classroom practice	Prior	2.39	1.15	18	3.13	1.12	24	-1.45, -0.03	-2.09**	40
	Current	3.05	1.27	19	3.28	1.14	25	-0.96, 0.51	-0.63	42
Feel evaluated fairly	Prior	3.78	1.11	18	3.88	1.04	24	-0.78, 0.58	-0.29	40
	Current	4.16	0.60	19	4.13	0.54	25	-0.32, 0.38	0.19	41
Feel last year will impact current year	Prior	2.11	1.13	18	2.79	1.22	24	-1.42, 0.06	-1.85*	40
	Current									

Note. Scale for Survey: Strongly Disagree 2- Disagree 3- Neither Agree nor Disagree 4- Agree 5- Strongly Agree
* = p< 0.1, **= p< 0.05, *** = p < 0.01

Interview

The data in Table 20 present the interview case for each code by subject area where the frequency is the number of interviews that included at least one occurrence of the code. This method was used rather than frequency of interviewees due to the small “n” of this data and the variable number of interviewees in each subject area. Each interviewee participated in two interviews for the study so the “n” reported equals twice the number of interviewees at a school. The percentage of occurrence for the interviews in each subject is also reported to allow for comparisons between the two groups. There were no stark differences between the two subject areas aside from a less frequent occurrence of statements related to external motivation appearing in English interviews versus Math interviews (27.8% versus 50%). Whether this difference in external motivation is related to the characteristics of the individual teachers interviewed for the study or to teachers of Math or English as a whole is unclear. It is also possible that these numbers are influenced by missing Math teachers from the interview sample at Riley and by only having one Math teacher participate at Phoenix.

Table 20

Occurrence of Codes in Interviews by Subject Area

	English (n= 18)		Math (n= 10)	
Codes	Frequency	Percentage	Frequency	Percentage
Motivation				
Internal	8	44.4%	5	50.0%
External	5	27.8%	5	50.0%
Observation Feedback	12	66.7%	8	80.0%
Negative	12	66.7%	8	80.0%
Positive	5	27.8%	4	40.0%
Testing Feedback	12	66.7%	8	80.0%
Negative	12	66.7%	8	80.0%
Positive	1	5.56%	1	10.0%
Work Decisions				
Strategy/How Taught	4	22.2%	4	40.0%
Curriculum/What Taught	4	22.2%	2	20.0%
Who is Taught	1	5.56%	1	10.0%
Response to Reform				
Acquiescence	6	33.3%	2	50.0%
Adaptation	4	22.2%	4	40.0%
Denial	2	11.1%	1	10.0%

Subject Area Specific Concerns

Observation concerns. The survey analysis findings are not surprising considering some of the specific concerns raised by teachers during the interview phase. It is possible that these findings relate to the background of the observer rather than the subject area of the observed teacher. As discussed in the previous chapter, Math teachers at Central were observed by a principal with a Math background, whereas English teachers at the same school often essentially evaluated themselves due to their assigned observer reportedly having difficulties operating the computer on which evaluation scores had to be recorded. In the case of one English teacher at Central, an observation of teaching was never conducted and instead she had been observed conducting a department meeting. Additionally, Math teachers at one other school, Charles, mentioned the value of having an administrator with a Math background observing lessons, which had occurred for those teachers in the previous study year.

Therefore, it is entirely possible that the experiences of teachers at those schools strongly affected the survey results where Math teachers felt that last year's evaluation would impact the study year and that observation feedback was used to modify classroom practice. Mr. Augustus explained how he imagined his administrator at Central, who had a Math background, could bring something different to the feedback given in a Math classroom versus another subject. While Mr. Augustus explained that the observations were "definitely focused on classroom management," he felt confident that when dealing with something very specific in the Math curriculum, that a non-Math person would be unable to appreciate the context and prior learning that students needed leading up to the lesson. However, Principal Nichols could provide feedback specific to the Math lesson he observed. Mr. Augustus was very sure of the Math competency of his evaluating administrator and explained, "I have had him, during observations, where he will see a kid that is struggling, and I am helping some other kids and he will go over and actually help him."

The description Mr. Augustus gives of his principal's feedback mirrors the experiences with coach observation that teachers found valuable, as discussed in Chapter 5. In that chapter, I described other informal forms of evaluation that teachers found valuable which included feedback from coach observations in cases where the teacher found the observer to be able to provide classroom relevant suggestions. The Math teachers at Charles in particular talked about their Math coach and highlighted that she was knowledgeable and had practical experience as a Math teacher, so she knew what it was like to teach a Math course.

Moreover, emerging literature supports the idea that the subject area background of the observer and the alignment between the background of an observer and the subject area observed may matter greatly in the types of feedback that a teacher receives (Bell et al., 2015). One of the assumptions behind the North Carolina teacher evaluation policy is that the rigorous standardization of the observation protocol would lead to more equitable observation experiences and better feedback for teachers. However, it seems that this feedback may be more useful when teachers are given the opportunity to receive feedback from an observer who not only understands a subject, but actively displays competence in the subject area.

Testing concerns. General concerns that teachers had about testing were addressed in Chapter 5, but two-subject specific concerns were raised in the interviews. English teachers had concerns that the tests which were administered to their students to gauge growth did not address all of the standards. On the other hand, Math teachers had a curriculum that had been in flux over the last several years and during the study year Math 2 teachers had to administer two tests to students, only one in which would be used to measure student growth or to provide feedback.

North Carolina adopted the Common Core Curriculum Standards (CCCS) for the 2012-2013 school year and the English/Language Arts (ELA) standards for high school, in the form most current at the time of this writing, have five anchor strands: Reading: Literature, Reading: Informational Text, Writing, Speaking and Listening, and Language. However, English teachers contend that the only strands addressed on the tests are the two Reading strands and some of the Language strand. When there are short answer writing portions on the test, teachers feel that those portions are not scored accurately because teachers receive the student score for the test on the same day it was administered. Teachers also cited examples of specific standards within the anchor strands that were known to not appear on the test and further described how the prioritizing of standards on the test influenced what they chose to teach in their classroom. For instance, many teachers chose to forego writing altogether in lower achieving classes in order to try and get students to pass the test and exhibit growth, whereas Listening/Speaking was not incorporated into lessons in a formal manner in any courses.

Such actions are similar to those demonstrated in research on gaming strategies for high-stakes testing that may result in a narrowing of curriculum to focus on tested aspects (Carnoy & Loeb, 2002; Ladd & Zelli, 2002; Rothstein & Mathis, 2013). Additionally, the different approaches teachers described in regard to changing curriculum between lower achieving classes and higher achieving classes could also represent a form of educational triage, except rather than removing students from the testing pool (Figlio, 2006; Figlio & Getzer, 2002) or diverting resources within a classroom (Booher-Jennings, 2005) teachers were adjusting the curriculum between classes in an attempt to get more students to pass the standardized test.

English teachers also raised the discrepancy between the idea of “College and Career Readiness” standards and an English test with content which was nearly a third based on poetry, which teachers contended indicated that the test was heavily grounded in literature despite the attempts to make ELA broader with the inclusion of the anchor standards. Mr. Allen described how the situation applied to students who wanted to go into technical trades such as welding or plumbing, “I think it is great that they should be exposed to poetry and that they should see what that has to say about society, but at the same time, is it fair that 30% of their final exam grade is based on a couple of random poems? When that has almost nothing to do with College and Career Readiness?” Similarly, Mrs. Williams stated that she had always joked that one year the tenth-grade test would include *War and Peace* and ask questions about Russian patronyms only to be horrified this year when she was proctoring the exam and noticed that *War and Peace* had been included. Mrs. Williams maintained that the literature and the concepts on the test were "not practical" to teach standard high school students.

In Math classes, the curriculum and standards have been shifting. Overall, the teachers seemed to agree with the shifts and felt like the state was responding to teacher concerns that the previous order of certain standards and topics did not align in a way that made sense across the four required Math courses. Teachers explained that certain courses used to be "heavy" in certain topics or that some concepts were presented out of order. However, during the study year, the standards for Math 2 changed significantly. So, some teachers in the study were expected to administer a statewide field test from which they would receive no data or feedback. A county developed exam was also administered and the student grades from that took about two weeks to receive because the data had to be transformed and managed at the county level. However, the district's position on this testing was reportedly to ask the Math teachers to lie and tell students that both tests, administered on separate days, would count toward student grades. By lying about the situation teachers were artificially creating pressure for students to perform in a situation absent of direct pressures to motivate students, a reform response described by Cohen (2011) that occurs in scenarios with high-stakes for teachers.

By the second interview, the teachers had already completed one round of Math 2 testing, but the dual tests were scheduled to be administered again at the end of the school year. The Math 2 teachers talked about how they "hated" lying to students and about how the two tests seemed like a waste of time and resources which did not provide any useful feedback. Teachers also described how the situation did not motivate them to have their students do well. Mr. Robbins summed up how the predicament presented an opportunity to game the system,

> [H]ow do you motivate kids? And …do I want the kids to bomb that field test? Because if every kid around the state just does horrible on it, then the state will do one of two things: they will either think that the test was too hard and they will write easier questions, which will benefit my kids in the future, or they will normalize it according to those awful grades, and it will mean that the curve in future years will be extremely low. One of those two things will happen, so you are not really motivating me to really push my kids to do extremely well on it.

Conclusion

Overall, some differences emerged between different groups for the three individual characteristics examined in this chapter. The results are similar when examining licensure and years of experience. For instance, both teachers who are provisionally licensed or have seven years or fewer of experience are more likely to report altering classroom practices due to evaluations when compared to teachers who are either professionally licensed or who have eight or more years of experience in North Carolina. Such differences could be due to many factors including a lack of guidance in how to model classroom practices, the more frequent occurrence of observations, or the higher stakes that evaluations carry for teachers with lower designations.

The differences in subject area may not be a result of a teacher's subject area background but instead a result of the contextual differences under which evaluation occurs. For instance, subject area results may have been influenced in part by a particularly poor observer in the English department at Central juxtaposed to a particularly competent observer for the Math department at the same school. In contrast, Math teachers at Central and at Charles both described positive experiences with having an observer with the same subject area background. So, feedback may be more useful when an observer is competent in the area observed and may contribute to changes in classroom practice in that manner. However, observers who are not competent in a subject provide feedback that has less value to a teacher.

Additionally, concerns about testing were raised by teachers of both subject areas. Specifically, English teachers were concerned that standards and entire anchor strands were left entirely off of the test. English teachers also expressed concern that the test was literature heavy and did not fit into the ideas of "College and Career Readiness" as espoused by the adopted standards. English teachers described some gaming behaviors such as narrowing of curriculum to focus on those standards and strands which were tested at the expense of non-tested elements of the curriculum. Meanwhile, Math teachers expressed specific concerns over the administration of a field test in Math 2. The concerns included that administering tests in this manner did not provide any reliable feedback and also led to opportunities to potentially game the system.

Overall, it appears that there are differences between provisionally and professionally licensed teachers and those with seven or fewer years of experience and those with more years of experience. These differences may be related to the higher stakes risk associated with evaluations for teachers without tenure protection or to the unique circumstances of being a newer teacher. Differences between subject areas in this study were not necessarily driven by inherently different characteristics between Math and English teachers, but rather by the unique circumstances under which each group taught and was evaluated. Additionally, there is some evidence to suggest that the subject area background may matter in regard to the value of feedback received and the extent to which such feedback can influence practice. Finally, conditions around testing led English teachers, and possibly Math teachers as well, to engage in gaming behaviors in an attempt to improve test scores.

CHAPTER 8: Conclusions and Implications

Prior teacher evaluation protocols were usually developed, or at least selected, at the local level and consisted of observations by school administrators. The results of such observations were usually bound within the school or district in which the observation was conducted, and principals relied on references to determine the potential ability of a new hire. Such systems were previously critiqued as rating too many teachers as high performing (Weisberg et al., 2009). Critics argued that a better system of teacher evaluation would be more standardized and include multiple measures to determine teacher proficiency, in turn more accurately gauging teacher competency by approaching proficiency from different angles and providing a better source of feedback to improve teacher practice (US Department of Education, 2009). While the critics of previous local-based systems presented valid points, their critiques emerged in a political atmosphere where student test scores had increasingly served as a proxy for student achievement, school-level governance was becoming increasingly consolidated at the state level, and teachers were increasingly portrayed by policymakers as individuals who had become complacent in their jobs under the safety of union strongholds. Federal initiatives such as RttT prompted a "rapid policy diffusion" of new evaluation policies which resulted in legislative changes in 46 states (Grissom & Youngs, 2015, p. 169). Mintrop and Sunderman termed the resulting legislation as the "third wave" of accountability where accountability for individual student success became narrowed to the focus of each individual teacher's impact (2013).

The teacher evaluation system used in North Carolina at the time of this dissertation was created in response to criticisms of prior local-based systems and in many ways, serves as an ideal example of what many policymakers felt an ideal evaluation system should look like. First, most educational policy, including teacher pay and graduation requirements, had been centralized at the state-level for several decades. Because of this, North Carolina already had pre-existing systems of technology which could be used to track student and teacher growth statewide under new accountability systems. Additionally, North Carolina was and is a "Right-to-Work State" and, thus, there has been limited job protection offered to teachers by teacher unions. These conditions allowed policymakers to elevate the stakes attached to teacher evaluations and connect observations and student test growth to the retention of employment, what Firestone had termed as "the most powerful incentive" to motivate teachers (2014, p. 102). Finally, North Carolina had already consolidated teacher evaluation at the state level with a rigorously standardized observation protocol prior to Race to the Top (RttT).

However, despite attempts to create a teacher evaluation system that answered the critiques of previous systems, the observation scores of teachers in the schools in this study were overwhelmingly rated proficient. For instance, 96.5% of the observation standards rated were marked proficient or higher in Charles, the lowest achieving school in this study which was 2.5% below the district average of 99%, indicating that there was a trend across Broadville County to rank teachers as proficient or higher. These results mirror other studies that demonstrate that the "widget effect" has persisted post evaluation reform (Sawchuck, 2013). Additionally, discrepancies existed between the two measures used to rank teachers: observation and student growth on standardized tests. For example, Phoenix teachers had 100% of their rated observation standards marked as proficient, yet had poor performance on the student growth standard with only 75% of teachers being rated as proficient. What has remained uncertain is how newer evaluation policies impact the work of teachers.

This dissertation examined the ways in which evaluation policy relates to teacher practice while considering various aspects of school and individual contexts. I then parsed out the ways in which school characteristics and individual-level characteristics may impact the evaluation-practice relationship. Though quantitative differences between schools were not found, there were qualitative differences in how evaluation was related to practice across sites. Differences were also found in the evaluation-practice relationship between teachers of different licensure levels and different levels of experience, possibly due to the increased risk evaluation carries to those in the lower designations. Finally, differences between the subject areas of Math and English were identified, but may have not been the result of unique characteristics of Math or English teachers. Rather, differences may have been influenced by the proficiency and approach of observers and a lack of subject area alignment between the observer and the classroom in English. In contrast, a subject area match was present for some of the Math teachers in the study. Therefore, it is important to examine the context of evaluation, particularly the capacity of the administration that conducts observation.

Despite attempts to standardize evaluation, there are factors that influence how observation is conducted in schools. For instance, the results of this study suggest that the characteristics and capacity of an observer do matter in how the observation protocol is interpreted and implemented. Additionally, the evaluation climate and culture, or evaluation scenario of a school, may also influence the ways in which teachers find evaluation motivating and how teachers approach feedback from evaluation. The results of this study provide insight into the relationship between teacher evaluation and classroom practice, an area that has previously been under researched despite the impact other high-stakes accountability policies have had on teaching practices and the teaching workforce.

Implications for Research

Teacher evaluation has gained much popularity as a research topic over the past decade. The importance of such work is amplified by the often drastic changes that occurred in state policies following RttT. Prior research focusing on the technical aspects of evaluation, including the potential issues of using both local-based observation tools and VAMs or student growth measures as part of evaluation, have been examined extensively (Baker et al., 2010; Bill and Melinda Gates Foundation, 2010; Corcoran, 2010; Glazerman et al., 2011; Goldhaber et al., 2013; Harris, 2009; Hill et al., 2001; McCaffrey et al., 2003; Rothstein & Mathis, 2013). For instance, the potential misidentification of teachers using VAMs has been extensively investigated (Goldhaber et al., 2013; Harris, 2009; Raudenbusch & Jean, 2012). Additionally, the infrastructure changes that accompany such systems have also been explored (Anagnostopoulos at al., 2013a; Mintrop & Sunderman, 2013; Thorn & Harris, 2013). However, at the time of writing, there was a gap in the scholarship examining the relationship between teacher evaluation policies and teacher practices. So, the work in this dissertation represents the next step in research on teacher evaluation policies, one in which the impacts of the policy on policy actors at the classroom level is examined.

This dissertation also represents part of the next generation of literature on how external accountability influences practice. Previous accountability policies have been examined for impacts on both the teaching workforce (Clotfelter et al., 2004) as well as on teacher practice (Carnoy & Loeb, 2002; Ladd & Zelli, 2002; Rothstein & Mathis, 2013). As accountability policy has now fully entered the “third wave” focused on individual-level accountability following RttT, it becomes important to revisit questions about how external pressures influence teachers. This is important because previous work on teacher responses to external pressures have demonstrated that teachers may engage in behaviors, in an attempt to meet policy demands, which carry financial or educational costs to schools. It is possible that such results are amplified when policy narrows to the level of individual accountability.

Additionally, this dissertation contributes to growing body of research on the relationship between evaluation and the context of evaluation, including observer background, observation protocols, and teacher characteristics, with particular attention to the capacity of the school leaders who are tasked with undertaking school-level evaluation. There is scant research available on how observers interact with observation protocols, yet newer work is emerging that examines the impact of the subject area background of observer upon both the ranking a teacher receives as well as the type of feedback received by the observed teacher (Bell et. al, 2015). Other work indicates that there may be differences between feedback received in different subject areas at the elementary-level (Burch & Spillane, 2003). Moreover, the frustrations expressed by teachers in this study mirrors recent work by Reinhorn et al. (2017), where teachers expressed disappointment with administrators lacking the background and experience to provide subject-specific recommendations for improvement. What is yet unclear is whether subject area differences, such as those observed in this study, stems from the nature of a subject or from the background and ability of the observer.

Implications for Policy and Practice

While teacher evaluation is often cited by policymakers as providing a source of feedback and motivation for teachers to improve classroom practice, the results of this study suggest that may not be true in some circumstances and contexts. The two assumptions that evaluation policy could simultaneously motivate teachers and provide feedback to improve practice did not play out as expected, at least in regard to the teachers in this study. So, is this failure for the policy to materialize as assumed a result of poor theory or poor implementation? Some teachers described other evaluation policies which were perceived as useful and well-implemented, particularly in regard to the use of instructional coaches. However, this was not a universal experience across teachers. For instance, teachers seemed to reject coaches who focused more on theory rather than actual teaching situations. Overall, teachers expressed that they felt that better feedback was received from an observer who knew the teacher's subject. Likewise, there were implementation issues of the formal evaluation policy, particularly in the case of Central, where one administrator did not conduct observation appropriately, whereas at Charles at least some of the teachers found the formal evaluation process to be helpful. It becomes apparent in looking at these scenarios across schools and across both formal and informal evaluation policies that teachers are likely to reject aspects of a policy which are perceived as coming from invalid sources.

Thus, it is important to understand the conditions under which policies are being implemented, particularly in regard to the capacity of the individual doing the evaluating. For instance, one reason teachers were critical of both components of the evaluation was due to the timing of feedback. For observations, teachers opined that the feedback was only about one class and that observations were often conducted in quick succession in a short amount of time or at the end of the year where improvements could not be implemented. These issues in timing are related to the capacity of leadership conducting the evaluations as well as limitations in resources (specifically time) to conduct the lengthy observation and feedback process. Testing feedback, on the other hand, was not available until the next year, which also prevented teachers from using the feedback to make meaningful change.

My study points to three implications for policy and practice. First, I explore the relationships that emerged in the data between leadership capacity and the success of the evaluation policy. Second, I describe how questions about the validity of the evaluation system in the constraints of the context of schools served as a barrier for evaluations being useful as a motivating tool or a feedback source. Finally, I describe how an evaluation system with conflicting messages about motivation in a high-stakes environment, as implemented in the schools in this dissertation, may serve to motivate school personnel to engage in undesirable behaviors.

Leadership and Evaluation

The observation instrument used in North Carolina at the time of this study was a lengthy, standardized, very detailed document that is meant to cover nearly every conceivable aspect of teaching. However, despite the detail of the observation instrument, the background, skill, preferences, and values of the human observer influenced how individual teachers were evaluated. For instance, in another study, factors such as teacher personality, philosophy, and effort were found to have contributed to evaluation ratings (Harris, Ingle, & Rutledge, 2014). Moreover, the standardization of the observation protocol used in the district in this dissertation could have been influenced through individual interpretations of the protocol or policy, by the evaluation scenario created by the climate and culture of the school, and by the observer's proficiency as an evaluator.

First, what an observer chooses to value in teaching may matter greatly in how the observation protocol gets interpreted, as was the case with Riley's the new administrator, Ms. Jefferson. The focal teachers reported that Ms. Jefferson tended to equate "good teaching" with the submission of lesson plans and expressed concern that this was impacting the way ratings were assigned in observation. According to the focal teachers, Ms. Jefferson placed emphasis on the completion of tasks rather than on what she saw happen in classrooms. The situation at Jefferson demonstrates that observers can choose to prioritize certain actions of teachers or interpret the observation instrument in a way that allows for such prioritization. Moreover, it is important to note that an overreliance on scores from evaluation, particularly as related to testing, may inhibit teacher autonomy which presents a challenge to intrinsic motivation (Firestone, 2014). Teachers at Phoenix described a “culture of improvement” that allowed for teacher autonomy and experimentation, which contrasts teachers at Riley who were apprehensive to even talk about evaluation with an outsider due to the administrator’s focus.

How the policy is interpreted by an observer also matters. Teachers across all of the school sites reported that the observation instrument was intended to be used as a growth instrument. Therefore, how a teacher is evaluated may hinge on their observing administrator’s interpretation of “growth.” For instance, teachers expressed that some administrators seemed to think that “growth” meant that a new teacher should always be ranked low regardless of past experience or of the performance observed in the classroom. The intentional lowering of initial evaluations in order to leave room for growth could be discouraging for a teacher who perhaps should have scored higher if ranked objectively and also is illustrative of how the scale used to rank teachers may not be truly standardized across all sites. The growth interpretation is one that emerges in this study, but there may be others. For instance, other studies have demonstrated that teachers respond to evaluation through the lens of their administration (Reinhorn et al., 2017) and that the persistence and strength of policy messages shapes the understanding and implementation of evaluation for administrators (Rigby, 2014).

Related to the first two points, evaluation scenarios also seem to be created within the school. Such scenarios appear to be primarily driven by administration and administrators’ individual approaches to evaluation, but also may reflect a long standing cultural tradition or climate component of the school as driven by the interpretation of evaluation policy. Three scenarios are presented in this dissertation, though there is certainly a possibility of more:

Technocratic, Autonomous/Self-Efficacious, and Consensus Lacking. The experiences teachers had with evaluation varied greatly depending on which scenario their school exhibited. As described previously, the technocratic scenario at Riley lead to feelings of duress in teachers whereas the lack of consensus at Central lead to some teachers finding evaluation to be invalid.

Finally, the proficiency of observers also interacts with observation protocols. The English teachers at Central described their observer, Mr. Reward, as an administrator who did not possess the skills or the proficiency needed to complete the observation instrument properly. Mr. Reward would often ask teachers to complete their own evaluation ratings and in one reported case, conducted an observation at an inappropriate time. Teachers at Central seemed to indicate that they did not forget when administrator "messed up" evaluation. Aside from issues with Mr. Reward, a teacher described how a previous assistant principal had made serious mistakes with overseeing testing which may have contributed to negative feelings toward evaluation.

While Central provides an extreme case of an observer lacking the proficiency to conduct observations, there is also some evidence from this study that suggests that the proficiency of an observer in the subject area being observed may also matter, particularly in regard to the quality of the feedback received. For instance, the Math teachers at Central and at Charles described situations under which they had been observed by an administrator with a Math background and described the feedback as useful and valid. Additionally, teachers who had positive experiences with curriculum coaches described a similar situation of receiving useful feedback that was directly relevant to their work in the classroom. Therefore, the impact an observer has on evaluation should be considered when observations are used as part of high-stakes decision making processes.

Perceptions of Validity

Teachers also expressed concerns over the validity of both components of the evaluation instrument. The concerns are crucial because in order for observations and test data to be used as a feedback source and a tool for motivation, teachers need to see the measure as valid. Many of the concerns around the validity of observations are related to the observer and are discussed above. For instance, teachers are unlikely to find feedback from an unskilled observer to be either useful or motivating. Teachers may also interpret evaluation differently depending on the evaluation scenario in their respective school or may approach feedback from an observation differently if they do not think the observer's focus is a valid component of teaching. However, teachers had other concerns about the validity of the observation instrument, such as questioning whether the frequency and timing of evaluations provides a good enough sample of their work to pass judgement of teaching ability. Similarly, teachers also expressed concerns that some of the standards on the observation may be too narrow for teachers to achieve every year based on a few observations.

Teachers also raised concerns over the testing component of evaluation, including that the test was very short and the questions did not address all of the standards in which teachers were tasked with teaching. Additionally, cut scores for students were very low, which teachers felt was misleading. Feedback on testing was described by one teacher as an "autopsy report" as it came much too late to be used to implement any classroom changes. Finally, the psychometric model which was used to calculate student growth and to evaluate teachers was difficult for teachers to understand. While teachers had been told about some components of the student growth equation, such as the removal of outliers which were meant to adjust for some of the very valid critiques researchers have presented on the use of VAMS in high-stakes situations (Harris, 2009 and others), teachers felt that the scores often seemed at odds with the realities of the classroom. Concerns over the validity of the evaluation components should be addressed if the purposes of evaluation are to include motivating teachers and providing feedback useful to improve teacher practice. Otherwise, it is important to provide opportunities to motivate and receive feedback in other ways from sources in which teachers do perceive validity.

Altered Teacher Behaviors

The results of this dissertation also suggest that the high-stakes evaluation system used in North Carolina at the time of the study may sometimes motivate teachers to engage in undesirable practices. While there were no extreme cases of gaming or cheating that emerged in this study, there was evidence that teachers sometimes altered practices to improve student test scores. For instance, teachers described how they may change the way in which questions are worded or the medium through which assignments are presented in order to familiarize students with formats found on the test. Teachers also cited examples of certain testing strategies they taught in order to assist students in becoming better test takers.

There were also examples of how curriculum was altered to meet evaluation requirements. In English, teachers chose to forgo certain standards because it was known they would not appear on the test. Teachers admitted that such practices meant that they did not successfully teach all the standards for their course. Additionally, English teachers stated that a narrowing of the curriculum was more common in the lower achieving courses where teachers felt more test prep would be necessary. This practice resulted in the withholding of certain parts of the curriculum from selected groups of students.

Teachers also cited some examples of changes in practice due to observation, such as being sure to incorporate technology on an observation day in order to ensure that standard would be met, but these examples were more benign than some of the alterations that were motivated by testing. Nonetheless, the examples that teachers shared about how they altered practice to accommodate either form of evaluation indicated that they were indeed motivated to make changes in order to receive a better score; however, some of the changes teachers were motivated to engage in may have unintended negative consequences for students.

Research also indicates that high-stakes accountability systems may result in increased turnover (Ingersoll, 2001). There were two instances of this that appeared in the interviews. Mr. Brown and Mr. Eagle were both teachers who had taught previously out of state, who were in their first years at their respective schools, and who both cited a focus on testing and pressure to have students perform well on tests as reasons for seeking other employment opportunities.

The behaviors noted above were most often reported by teachers who were provisionally licensed and/or had seven or fewer years of experience in North Carolina and were therefore subjected to increased observations and were unprotected by career status. It's unclear whether teachers with provisional licensure or seven years or fewer designations reported engaging in these behaviors more frequently because they were less experienced teachers or because evaluation held higher stakes for them than more experienced teachers. Summatively, these examples of teacher behavior suggest that teachers do respond to high-stakes evaluation in a manner similar to studies done on teacher response to other accountability measures (Rothstein & Mathis, 2013; and others). Therefore, the benefits of high-stakes evaluation policy should be weighed against these unintended consequences.

Reconciling Evaluation Policy for Both High and Low Stakes Purposes

The evaluation policy in North Carolina as well as elsewhere in the country is partially meant to serve as a tool to regulate the quality of teachers in the classroom. Previous research has demonstrated that while principals do report using evaluation to move poorly performing teachers towards dismissal, such teachers often leave before formal dismissal can occur (Kraft & Gilmour, 2017). And while the use of multiple measures has attempted to mitigate previous concerns over the use of local observations, when making human resources decisions principals have reported relying more on observations, which are perceived as more specific and transparent, than on VAMs or test scores which are not timely and are opaque (Goldring et al., 2015). One study suggests that the perspective of school administrators is that effective teaching is broader than what can be expressed in test scores and, as also demonstrated in this dissertation, such interpretations are subject to a principal's prior knowledge, connection with the policy message, and the social context of the school (Rigby, 2014). Similarly, VAMs have been shown to correlate with principal assessments of a teacher's ability to raise test scores, but not with other aspects of teaching, making VAMs a narrow predictor of a teacher's ability to do their job (Grissom, Loeb, & Doss, 2016). Moreover, the high-stakes nature of current evaluation policy may make it difficult for administrators to honestly assess their teachers, particularly when replacing a teacher may be difficult or when administrators feel like they lack the capacity to effectively evaluate in a high-stakes scenario. For instance, a study by Grissom and Loeb (2017) found that principals tended to evaluate more positively on higher stakes evaluations that on low stakes. This provides a possible explanation for why teachers still tend to be highly rated by administrators in the schools in this study and elsewhere.

So, can evaluation be simultaneously a formative feedback experience and a summative high-stakes tool for human resource decisions? As far back as 1988, Popham referred to the "dysfunctional marriage" between the two concepts and Firestone (2014) outlined how those concepts involved two competing theories of motivation that stymied progress. However, there are some ways in which this relationship could be improved under current policy.

If the policy is to use evaluation for feedback, then some changes must be made to make current systems more effective. Critiques that evaluation instruments are too broad could be addressed by instead providing focused feedback on a few targets. Teachers, such as those in this dissertation, may perceive that it is unfair that administrators make judgements on areas of practice where an evidence-based recommendation for improvement cannot be provided. Henry and Guthrie (2016) explained that "in a system where everything is a priority, nothing is a priority" (p. 153). Teachers are unable to improve if they are unsure of what needs improving. Likewise, principals need the training to provide specific, actionable feedback that will be of use to teachers.

Likewise, if teachers are to be evaluated using VAMs than the timeline for the return of scores and feedback should be shortened and shared with teachers in a way that would allow evaluators to make meaningful changes to their practice immediately. If the timeline for providing feedback cannot be tightened and/or if the feedback provided cannot be made to be more specific and useful, then observers need to be able provide feedback that will allow teachers to improve their practice and thus improve test scores (Henry & Guthrie, 2016). This would involve additional training for observers and the creation of professional opportunities for teachers to examine and interpret the data. Additionally, there are several research supported school-level supports which can be provided to help teachers use feedback including: relevant professional development opportunities, timely and specific feedback tied to effective teaching, and the influence of collegial relationships (Sun, Penuel. Frank, Gallagher, & Youngs, 2013). In this dissertation, teachers reported successful informal evaluation experiences when certain conditions were met. For instance, teachers reported successful experiences when they had common time to work together on local assessments, whether developed at the district or department level. Overall, research has demonstrated that teachers working in more supportive environments are more likely to improve their effectiveness over time (Kraft & Papay, 2014). These opportunities for professional support need to be created in concert with the formal evaluations.

There may also be promise in the inclusion of subject specific observation protocols. The use of generic instruments, such as the one in this study, assume that the same types of knowledge and practices are suitable across all grade levels and subject areas while simultaneously assuming that evaluators can assess instruction in areas where they do not have background (Youngs & Whittaker, 2016). Additionally, commercially available subject specific protocols tend to focus on lessons as opposed to other areas of teaching, such as the use of summative assessments and data analysis ability (Young & Whittaker, 2016). It may be that principal observations should be combined with other types of observations with different foci to gain a more well-rounded impression of teacher ability. In this dissertation, teachers who did not find validation in their principal's assessments often cited other sources, such as curriculum coaches or colleagues who addressed lesson and classroom specific aspects of teaching, as testaments to personal skill and sources of valuable feedback.

Similarly, if the evaluation policy is to attach high-stakes to teacher evaluation, then the bias that was a focus of critiques of previous locally developed systems could be addressed by designing a training system which utilizes a calibration technique and multiple observers (Youngs & Grissom, 2016). The use of multiple observers could include individuals who have expertise in the teacher's subject area. Additionally, given criticisms of such uses in current research, stronger evidence of the validity and reliability of current evaluation systems for making high-stakes human resource decisions should be presented so that both teachers and administrators can be more confident in the accuracy of ratings (Youngs & Grissom, 2016). This may allow principals to feel that they can give more honest critiques of teachers in observations rather than distributing high ratings across the workforce. The teachers who participated in this dissertation showed great distrust of the accuracy of both the observation and testing components of the evaluation. Such distrust could be mitigated by changes to the evaluation process.

Limitations

There are three main limitations to this study. First, this study is bound by the specific context in which the study schools are situated. While a variety of schools were deliberately sampled for this study, all of the schools are located in the same school system. The policy of interest is a state-level policy; however, it is unclear what, if any, influence district-level priorities and initiatives, or the physical location of the county examined here in proximity to other counties and states, may have had on the relationship between evaluation and practice. Additionally, the policy investigated here is unique to the state of North Carolina. While North Carolina serves as a model of many of the tenants espoused by the RttT application requirements and while many states have adopted legislation that is similar to North Carolina's as a result of RttT, no other state will have the same policy history, concurrent policies, and cultural, social, and historical identities that North Carolina does.

The context of North Carolina is one of higher stakes than other states where local unions may be stronger. For instance, there is great variability in how states implemented RttT inspired teacher evaluation policies and often districts are able to select local models. However, in North Carolina, teachers are evaluated under the same system and evaluation ratings are electronically recorded at the state level, which may impact a teacher's future career prospects or ability to be mobile across the state. Furthermore, the North Carolina evaluation model is a growth model and teachers are expected to exhibit continuing growth, which may result in unintended consequences such as initially receiving lower ratings or the inflation of ratings among more experienced teachers.

Additionally, this study only examined high school teachers of two subject areas: Math and English. Therefore, the dissertation does not address other grade levels or subject areas which may have very different experiences and perspectives from high school Math and English teachers. While this study provides important information on the relationship between teacher evaluation and teacher practice in a state with a high-stakes, statewide teacher evaluation policy, it is unclear whether the results would be replicated elsewhere or under different circumstances or with different populations of teachers.

A second limitation of this study is the assessment of differences in Evaluation Conditions and School Evaluation Effectiveness. An initial goal of this dissertation was to identify ways in which classroom practice was differentially impacted by sampling schools of various conditions and effectiveness levels. However, despite differences in scores and differences which emerged in the qualitative work, no statistically significant differences occurred between teachers at different schools on the survey measures. It may be that measuring evaluation conditions and effectiveness in a different way may have yielded different quantitative results. It is also possible that a finer grained analysis may have been necessary to discern differences in conditions. For instance, the Math and English teachers at Central had very different experiences with evaluation. So, examining conditions at a department level may have produced different results. It is also possible that any potential results were understated by the small sample sizes utilized in this study. What is clear from this study is that evaluation policy, and particularly the observation component of evaluation, was implemented very differently across the four school contexts despite the rigorous standardization of the protocol at the state level.

Finally, a second focus of this study was to discern if there were differences between Math and English teachers and the relationship between evaluation and practice. While differences were found statistically, the qualitative work revealed differences in how teachers of these subject areas were observed as well as issues with tests which were specific to each subject area. Therefore, I was unable to determine if the subject area differences were inherent characteristic of either Math or English teachers or instead a result of the unique conditions under which teachers of each subject were evaluated. Stronger conclusions may have been drawn from a larger sample, or at least, from a more even sample of teachers. No Math teachers from Riley agreed to be interviewed for this project, which the department chair stated was presumably because of the school's new administrator's increased focus on observation. Additionally, only one of the Math teachers at Phoenix was available for interview. This created unbalance in the sample as well as a lack of representation in the focal interviews for one entire Math department from a study school. This limitation does not mean that subject area does not matter, but rather that the context in which a teacher of a certain subject area is evaluated may matter more than the subject itself.

Concluding Thoughts

From a policy perspective, it is important to consider that contextual differences exist in schools and so formal evaluation may not always be a useful source of feedback to teachers and may not accurately reflect an individual's teaching abilities. Currently, such evaluations are high-stakes and are attached to teacher job retention policies. Evaluation results are also reported to the state-level and follow a teacher throughout their career, leading to the possibility that a teacher who has had inaccurate but poor evaluations may be negatively impacted in the future. Therefore, it is important to consider the potential benefits of using an imperfect evaluation instrument, whether that instrument is observation, student growth, or a combination of both, against potential individual-level costs.

If evaluation is to serve as both a motivator to improve classroom practice and as a source of feedback to teachers, then certain conditions of the evaluation may need to be changed. Qualitative results suggest that ongoing formative feedback by an observer or by multiple observers who can identify what good teaching looks like in a context is more valuable and motivating to teachers than a summative assessment. Additionally, the high-stakes nature of current evaluation policy may drive teachers to engage in practices which may be detrimental to student learning. Moreover, these practices may actually be rewarded under current evaluation systems. Additionally, such detrimental practices may be amplified in teachers who undergo more frequent evaluations without career protections. This concern is particularly relevant for North Carolina, as the legislation which is current at the time of writing is effectively phasing out teacher career status for all teachers.

The assumptions behind teacher evaluation policy requires the policy to be both high-stakes and be used to weed-out low performing teachers while simultaneously providing the type of feedback that can support a teacher’s development on a growth model evaluation. Firestone (2014) had argued that the success of the type of evaluation system seen in North Carolina, which focuses on the use of both external and internal motivating factors, is stymied by the inherent conflicts between the two theories of motivation. The results of this dissertation support this hypothesis, though there are several areas where the system could be improved to better accommodate both goals. The policy examined in this study is problematic because it tasks administrators with conducting high-stakes evaluations and providing formative feedback to all teachers up to four times a year. Yet, principals often lack the training and time resources to evaluate teachers in a high-stakes manner and to simultaneously provide constructive feedback to allow for systematic improvements. To accomplish a better balance, evaluation would need to be lower stakes, more formative, and focus on all teachers, not just a concentration of newer teachers.

At the time of this writing, there is a gap in the literature on how formal teacher evaluation policy is related to classroom practice. This is an important question to consider because evaluation, by definition, defines what is valued in whatever is being appraised. Additionally, such policies are touted by policymakers as being necessary to motivate teachers to do better jobs and to provide feedback for them to do so. Therefore, it is important to consider whether or not the formal policies do motivate and provide feedback to teachers and if such policies do these things than to consider in what ways teacher practice changes as a result? While questions around the evaluation and practice relationship could benefit from future work using larger sample sizes and perhaps spanning additional levels of schooling and different subject areas, this dissertation begins to answer important questions around evaluation and practice as related to the study context. Such information is useful when weighing the costs and benefits of high-stakes teacher evaluation policies.

APPENDIX

APPENDIX A: Survey Instrument

Part I: Demographic Questions (Short Answer)

1.) Including this year, how many years have you been teaching?

2.) Including this year, how many years have you been teaching in North Carolina?

3.) Including this year, how many years have you been teaching at this school?

4.) What is your certification level? Provisional, Professional/Career, Other

5.) What subjects are you certified in?

6.) What grades are you certified to teach?

7.) Have you ever taught a course that was assessed by an End of Course (EOC) or End of Grade (EOG) exam?

8.) Did you teach English II, Math I, or Biology last year?

9.) Are you teaching English II, Math I, or Biology this year?

10.) Indicate your level of agreement with the following questions about the conditions in this school: (Scale: Strongly Disagree, Disagree, Agree, Strongly Agree, Don't Know)

a. Teachers are held to high professional standards for delivering instruction.

b. Teacher performance is assessed objectively

c. Teachers receive feedback that can help them improve teaching.

d. The procedures for teacher evaluation are consistent

e. State assessment data are available in time to impact instructional practices.

f. Local assessment data are available in time to impact instructional practices.

g. Teachers use assessment data to inform their instruction.

h. State assessment data are available in time to impact instructional practices.

i. State assessments provide schools with data that can help improve teaching.

j. State assessments accurately gauge students' understanding of standards.

Part II: Prior Year

Indicate your level of agreement with the following questions about your practices in the classroom from the following year (2015-2016) and (B) current year (2016-2017). (Scale: Strongly Disagree, Disagree, Neither Agree or Disagree, Agree, Strongly Agree, Not Applicable)

11.) Last year, I modified classroom practice in anticipation of an upcoming evaluation.

12.) Last year, I modified classroom practice using feedback from my evaluation.

13.) Last year, I was concerned that my evaluation results could impact future employment.

14.) Last year, I was concerned that my evaluation may label me as a bad teacher.

15.) Last year, I was concerned that my evaluation does not accurately reflect my competency as a teacher.

16.) Last year, I chose curriculum based on what I will be evaluated on.

17.) Last year, I chose teaching strategies based on what I will be evaluated on.

18.) Last year, I directed focus on certain students based on what I will be evaluated on.

19.) Last year, I used test data to modify classroom practice.

20.) Last year, I used observation feedback to modify classroom practice.

21.) I felt I was evaluated fairly in the previous school year.

22.) Last year's evaluation will impact my decisions about classroom practice in the new school year.

Part III: Current Year

Indicate your level of agreement with the following questions about your practices in the classroom from the current year (2016-2017). (Scale: Strongly Disagree, Disagree, Neither Agree or Disagree, Agree, Strongly Agree, Not Applicable)

23.) This year, I will modify classroom practice in anticipation of an upcoming evaluation.

24.) This year, I will modify classroom practice using feedback from my evaluation.

25.) This year, I am concerned that my evaluation results could impact future employment.

26.) This year, I am concerned that my evaluation may label me as a bad teacher.

27.) This year, I am concerned that my evaluation does not accurately reflect my competency as a teacher.

28.) This year, I will choose curriculum based on what I will be evaluated on.

29.) This year, I will choose teaching strategies based on what I will be evaluated on.

30.) This year, I will direct focus on certain students based on what I will be evaluated on.

31.) This year, I will use test data to modify classroom practice.

32.) I will use observation feedback to modify classroom practice.

33.) I feel I will be fairly during this school year.

REFERENCES

REFERENCES

Allen, J. P., Pianta, R. C., Gregory, A., Mikami, A. Y., & Lun, J. (2011). An interaction-based approach to enhancing secondary school instruction and student achievement. *Science, 333*(6045), 1034–1037.

Anagnostopoulos, D., Rutledge, S. R., & Jacobsen, R. (2013a). Mapping the information infrastructure of accountability. In D. Anagnostopoulos, S. Rutledge, & R. Jacobsen (Eds.), *The infrastructure of accountability* (pp. 1-21). Cambridge, MA: Harvard University Press.

Anagnostopoulos, D., Rutledge, S. R., & Jacobsen, R. (2013b). The infrastructure of accountability: Tensions, implications, and concluding thoughts. In D. Anagnostopoulos, S. R. Rutledge, & R. Jacobsen (Eds.), *The infrastructure of accountability* (pp. 213-228). Cambridge, MA: Harvard University Press.

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., & Shepard, L. A. (2010). *Problems with the use of student test scores to evaluate teachers.* Washington, DC: Economic Policy Institute.

Ball, D. L., & Bass, H. (2000). Interweaving content and pedagogy in teaching and learning to teach: Knowing and using mathematics in J. Boaler (Ed.), *Multiple perspectives on the teaching and learning of mathematics* (pp. 83-104). Westport, CT: Ablex.

Ballou, D., & Podgursky, M. (1998). Teacher recruitment and retention in public and private schools. *Journal of Policy Analysis and Management*, *17*(3), 393-417.

Bandura, A. (1997). *Self-efficacy in changing societies.* New York, NY: Cambridge University Press.

Barnes, C. R. (2011). “Race to the Top” only benefits big government. *Journal of Law and Education, 40*(2), 393-402.

Bell, C., Drake, C., Wilson, M., Frasier, A., Qi, Y., McCaffrey, D., Lockwood, J. R., & Kim, J. (2015). *Subject specific and general observation protocols as tools for the evaluation and improvement of teaching*. Paper session presented at the meeting of the American Education Research Association. Chicago, IL.

Berg, B. L. (2007). *Qualitative research methods for the social sciences* (6th Edition). San Francisco, CA: Pearson Education.

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

Bill & Melinda Gates Foundation. (2013). *Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET Project's three-year study.* Seattle, WA: Author.

Black, P., & William, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation, and Accountability, 21*(1), 5-31.

Booher-Jennings, J. (2005). Below the bubble: "Educational Triage" and the Texas accountability system. *American Educational Research Journal, 42*(2), 231-268.

Boyd, D. J., Grossman, P., Lankford, H., Loeb, S., Wyckoff, J. (2006). How changes in entry requirements alter the teacher workforce and affect student achievement. *Education Finance and Policy*, *1*(2), 176–216.

Boyd, D. J., Grossman, P. L., Lankford, H., Loeb, S., & Wyckoff, J. (2009). Teacher preparation and student achievement. *Educational Evaluation and Policy Analysis*, *31*(4), 416-440.

Boyd, D., Lankford, H., Loeb, S., Rockoff, J., & Wyckoff, J. (2008). The narrowing gap in New York City teacher qualifications and its implications for student achievement in high-poverty schools. *Journal of Policy Analysis and Management, 27,* 793–818.

Bryk, A., Sebring, P. B., Allensworth, E., Luppescu, S., & Easton, J. (2010). *Organizing schools for improvement: Lessons from Chicago*. Chicago, IL: University of Chicago Press.

Burch, P., & Spillane, J. P. (2003). Elementary school leadership strategies and subject matter: Reforming mathematics and literacy instruction. *The Elementary School Journal*, *103*(5), 519-535.

Carnoy, M., & Loeb, S. (2002). Does external accountability affect student outcomes? A cross-state analysis. *Educational Evaluation and Policy Analysis*, *24*(4), 305-331.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2011). *The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood*. NBER Working Paper Series. Working Paper. National Bureau of Economic Research. Cambridge, MA. Retrieved from http://obs.rc.fas.harvard.edu/chetty/value_added.pdf

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2007). Teacher credentials and student achievement: Longitudinal analysis with student fixed effect. *Economics of Education Review, 26*, 673–682.

Clotfelter, C. T., Ladd, H. F., Vigdor, J. L., Diaz, R. A. (2004). Do school accountability systems make it more difficult for low-performing schools to attract and retain high-quality teachers? *Journal of Policy Analysis and Management*, *23*(2), 251-271.

Coburn, C. E. (2004). Beyond decoupling: Rethinking the relationship between the institutional environment and the classroom. *Sociology of Education*, *77*(3), 211-244.

Cohen, D. K. (1990). A revolution in one classroom: The case of Mrs. Oublier. *Educational Evaluation and Policy Analysis*, *12*(3), 311-329.

Cohen, D. K., Raudenbusch, S. W., & Ball, D. L. (2003). Resources, instruction, and research. *Educational Evaluation and Policy Analysis, 25*(2), 119-142.

Cohen, D. K. (2011). *Teaching and its predicaments.* Cambridge, MA: Harvard University Press.

Cohen-Vogel, L. (2011). Staffing to the test: Are today's school personnel practices evidence-based? *Educational Evaluation and Policy Analysis, 33*(4), 483- 505.

Conley, D. T., Drummond, K. V., Gonzalez, A., Seburn, M., Stout, O., and Rooseboom, J. (2011). Lining up: The relationship between the Common Core State Standards and five sets of comparison standards. Educational Policy Improvement Center. Retrieved from www.epiconline.org

Corcoran, S. P. (2010) *Can Teachers be Measured by Test Scores? Should They Be? The Use of Value-Added Measures of Teacher Effectiveness in Policy and Practice.* Annenberg Institute for School Reform.

Cusick, P. (1983). *The egalitarian ideal and the American high school.* New York, NY: Longman.

Darling-Hammond, L., Holtzman, D. J., Gatlin, S. J., & Heilig, J. V. (2005). Does teacher preparation matter? Evidence about teacher certification, Teach for America, and teacher effectiveness. *Education Policy Analysis Archives*, *13*(42).

Deci, E. L. & Ryan, R. M. (1996). Need satisfaction and the self-regulation of learning. *Learning & Individual Differences, 8*(3), 165-184.

Deyhle, D. L., Hess, G. A., & LeCompte, M. D. (1992). Approaching ethical issues for qualitative researchers in education. In M. LeCompte, W. Millroy, & J. Preissle (Eds.), *Handbook of qualitative research in education* (pp. 598–641). San Diego: Academic Press.

Derrington, M. L., & Campbell, J. W. (2013). The changing conditions of instructional leadership: Principals' perceptions of teacher evaluation accountability measures. In B. Barnett, A. R. Shoho, & A. J. Bowers (Eds.), *School and district leadership in an era of accountability* (pp. 231-251). Charles, NC: Information Age Publishing.

Educator Effectiveness Database. (2015). *Information on educator effectiveness data.* Retrieved from http://apps.schools.nc.gov/ords/f?p=155:1

Erpenbach, W. J. (2014). *A study of states' requests for waivers from requirements of the No Child Left Behind Act of 2001: New developments in 2013-2014.* Washington, DC: Council of Chief State School Officers.

Figlio, D. N. (2006). Testing, crime and punishment. *Journal of Public Economics*, *90*(4), 837-851.

Figlio, D. N., & Getzler, L. S. (2002). *Accountability, ability and disability: Gaming the system* (No. w9307). National Bureau of Economic Research.

Figlio, D. N., & Winicki, J. (2005). Food for thought: the effects of school accountability plans on school nutrition. *Journal of Public Economics*, *89*(2), 381-394.

Firestone, W. A. (2014). Teacher evaluation policy and conflicting theories of motivation. *Educational Researcher*.

Firestone, W. A. & Pennell, J. R. (1993). Teacher commitment, working conditions, and differential incentive policies. *Review of Educational Research, 63*(4), 489-529.

Firestone, W. A., & Rosenblum, S. (1988). Building commitment in urban high schools. *Educational Evaluation and Policy Analysis, 23*(4), 285-300.

Floden, R. E., Porter, A. C., Schmidt, W. H., & Freeman, D. J. (1980). Don’t they all measure the same thing?: Consequences of standardized test selection. In E. L. Baker and E. S. Quellmalz (Eds.) *Educational testing and evaluation, design, analysis, and policy* (pp. 109-120). Beverly Hills, CA: Sage.

Garda, R. A., & Doty, D. S. (2013). The legal impact of emerging governance models on public education and its office holders. *Urban Lawyer, 45*(21).

Garet, M. S., Porter, A. C., Desimone, L., Birman, B. F., & Yoon, K. S. (2001). What makes professional development effective? Results from a national sample of teachers. *American Educational Research Journal, 38*(4), 915-946.

Giroux, H. (1985). Teachers as transformatory intellectuals. *Critical Educators*, 46-49.

Glazerman, S., Goldhaber, D., Loeb, S., Raudenbusch, S., Staiger, D., & Whitehurst, G. J. (2011). *Passing muster: Evaluating teacher evaluation systems.* Washington, DC: The Brookings Brown Center Task Group on Teacher Quality.

Glesne, C. (2006). *Becoming qualitative researchers: An introduction (3rd ed.).* Boston: Pearson.

Goldhaber, D. D., & Brewer, D. J. (1997). Why don't schools and teachers seem to matter? Assessing the impact of unobservables on educational productivity. *Journal of Human Resources*, *Summer*, 505-523.

Goldhaber, D., & Hansen, M. (2010). Implicit measurement of teacher quality: Using performance on the job to inform teacher tenure decisions. *American Economic Review, 100*(2), 250-255.

Goldhaber, D., Goldschimdt, P., & Tseng, F. (2013). Teacher value-added at the high school level: Different models, different answers? *Education Evaluation and Policy Analysis.* Retrieved form http://epa.sagepub.com/content/early/2013/01/15/0162373712466938

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

Grissom, J. A., Kalogrides, D., & Loeb, S. (2012). Strategic Staffing: Examining the Class Assignments of Teachers and Students in Tested and Untested Grades and Subjects. Paper presented at the annual meeting of the Association of Education Finance and Policy.

Grissom, J. A. & Loeb, S. (2017). Assessing principals' assessments: Subjective evaluations of teacher effectiveness in low- and high-stakes environments. *Education Finance and Policy*, *12*(3), 369-395.

Grissom, J. A. & Youngs, P. (2015). *Improving teacher evaluation systems: Making the most of multiple measures*. New York, NY: Teachers College Press.

Grissom, J. A., Loeb, S., & Doss, C. (2016). The multiple dimensions of teacher quality: Does value-added capture teachers' nonachievement contributions to their schools? In Grissom, J. A. & Youngs, P. (Eds.), *Improving teacher evaluation systems: Making the most of multiple measures*. (pp. 37-50). New York, NY: Teachers College Press.

Hackman, J. R., & Oldham, G. R. (1980). *Word redesign.* Reading, MA: Addison-Wesley.

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

Hanushek, E. A., & Rivkin, S. G. (2010) Generalizations about using value-added measures of teacher quality. *American Economic Review, 100*(2), 267-271.

Harris, D. N. (2009). Would accountability based on teacher value added be smart policy? An examination of the statistical properties and policy alternatives. *Education Finance and Policy*, *4*(4), 319-350.

Harris, D. N., Ingle, W. K., & Rutledge, S. A. (2014). How teacher evaluation methods matters for accountability: A comparative analysis of teacher effectiveness ratings by principals and teacher value-added measures. *American Educational Research Journal, 5*(1), 73–112.

Hart, A. W., & Murphy, M. J. (1990). New teachers react to redesigned teacher work. *American Journal of Education, 98,* 224-250.

Haycock, K. (1998) Good teaching matters. *Educational Trust: Thinking K-16.* Retrieved from http://edtrust.org/sites/edtrust.org/files/public_actions/files/k16_summer98.pdf

Henry, G. T. & Guthrie, J. E. (2016). Using multiple measures for developmental teacher evaluation. In Grissom, J. A. & Youngs, P. (Eds.), *Improving teacher evaluation systems: Making the most of multiple measures*. (pp. 143-155). New York, NY: Teachers College Press.

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

Ingersoll, R. M. (2001). Teacher turnover and teacher shortages: An organizational analysis. *American Educational Research Journal*, *38*(3), 499-534.

Ingersoll, R. M., & May, H. (2012). The magnitude, destinations, and determinants of mathematics and science teacher turnover. *Educations Evaluation and Policy Analysis, 34*(4), 435-464.

Jacob, B. A., & Levitt, S. D. (2003). *Rotten apples: An investigation of the prevalence and predictors of teacher cheating* (No. w9413). National Bureau of Economic Research.

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher*, *33*(7), 14-26.

Johnson, S. M., & Birkland, S. E. (2003). Pursuing a “sense of success”: New teachers explain their career decisions. *American Educational Research Journal, 40*(3), 581-617.

Kennedy, M. M. (2005). *Inside teaching: How classroom life undermines reform*. Cambridge, MA: Harvard University Press.

Kraft, M. A. & Gilmour, A. (2016). Can principals promote teacher development as evaluators? A case study of principals’ views and experiences. *Educational Administration Quarterly, 52*, 711-753.

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

Kushman, J. W. (1992). The organizational dynamics of teacher workplace commitment: A study of urban elementary and middle schools. *Educational Administration Quarterly*, *28*(1), 5-42.

Ladd, H. F., & Zelli, A. (2002). School-based accountability in North Carolina: The responses of school principals. *Educational Administration Quarterly*, *38*(4), 494-529.

Lipsky, M. (2010). *Street-level bureaucracy: Dilemmas of the individual in public service* (30th anniversary expanded ed.). New York, NY: Russell Sage Foundation.

Loeb, S., Kalogrides, D., & Beteille, T. (2012). Effective schools: Teacher hiring, assignment, development, and retention. *Education Finance and Policy, 7,* 269–304.

Louis, K., Dretzke, B., & Wahlstrom, K. (2010). How does leadership affect student achievement? Results from a national US survey. *School Effectiveness and School Improvement, 21,* 315-336.

Mann, H. (1868). *Life and works of Horace Mann* (Vol. 3). Walker, Fuller and Company.

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M. & Hamilton, L. S. (2003). *Evaluating value-added models for teacher accountability.* Santa Monica, CA: RAND.

McDonell, L. M. (2005). No Child Left Behind and the federal role in education: Evolution or revolution? *Peabody Journal of Education*, *80*(2), 19–38.

Miles, M. B., Huberman, A. M., & Saldana, J. (2014). Qualitative data analysis: A method sourcebook. *CA, US: Sage Publications*.

Milgrom, P. R., & Roberts, J. D. (1992). *Economics, organization and management.* Englewood Cliffs, NJ: Prentice-Hall.

Mintrop, H., & Sunderman, G. L. (2013). The paradoxes of data-driven school reform: Learning from two generations of centralized accountability systems in the United States. In D. Anagnostopoulos, S. R. Rutledge, & R. Jacobsen (Eds.), *The infrastructure of accountability*. (pp. 23-40). Cambridge, MA: Harvard University Press.

Neal, D., & Schanzenbach, D. W. (2010). Left behind by design: Proficiency counts and test-based accountability. *The Review of Economics and Statistics, 92*(2), 263-283.

Newman, A. (2013). *Realizing educational rights: Advancing school reform through courts and communities.* Chicago, IL: University of Chicago Press.

North Carolina Board of Education. (2012). *ESEA flexibility request.* Retrieved from http://www2.ed.gov/policy/eseaflex/approved-requests/nc.pdf

North Carolina Board of Education. (2012). *16 NCAC 6C .0504. policy on standards and criteria for evaluation of professional school employees.* Retrieved from http://sbepolicy.dpi.state.nc.us/policies/TCP-C-006.asp

North Carolina State Board of Education. (2012). *115C-333; N.C. Constitution, Article IX, Sec.5. Teacher evaluation process.* Retrieved from http://ncrules.state.nc.us/ncac/title%2016%20%20education/chapter%2006%20%20 elementary%20and%20secondary%20education/subchapter%20c/16%20ncac% 2006c%20.05 03.pdf

North Carolina Department of Public Instruction. *North Carolina Educator Effectiveness Data.*

Retrieved from http://apps.schools.nc.gov/pls/apex/f?p=155:1

North Carolina Teacher Working Condition Survey. (2017). *Reports from TWC 2016.* Retrieved from https://ncteachingconditions.org/results

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

Oliver, C. (1991). Strategic responses to institutional processes. *Academy of Management Review*, *16*(1), 145-179.

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

Porter, A., McMaken, J., Hwang, J. & Yang, R. (2011). Common Core Standards: The new U.S. intended curriculum. *Educational Researcher 40*(103). 103-116.

Porter, A. C., Polikoff, M. S., & Smithson, J. (2009). Is there a defacto national intended curriculum? Evidence from state content standards. *Educational Evaluation and Policy Analysis. 31*(3). 238-268.

Powell, A., Farrar, E., & Cohen, D. (1985). *The shopping mall high school.* Boston, MA: Houghton Mifflin.

Powell, E. (2013). The quest for teacher quality: Early lessons from Race to the Top and state legislative efforts regarding teacher evaluation. *DePaul Law Review 62*(1061).

Raudenbusch, S. W., & Jean, M. (2012). How should educators interpret value-added scores? Retrieved from htpp://carnegieknowledgenetwork.org/briefs/value-added/interpreting-value-added

Reinhorn, S. K., Johnson, S. M. & Simon, N. S. (2017) Investing in development: Six high-performing, high-poverty schools implement the Massachusetts Teacher Evaluation Policy. *Educational Evaluation and Policy Analysis, 39*(3), 383-406.

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

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review, 94,* 247–25.

Rockoff, J. E., Jacob, B. A., Kane, T. J., & Staiger, D. O. (2011). Can you recognize an effective teacher when you recruit one? *Education Finance and Policy, 6,* 43–74.

Rothstein, J., & Mathis, W. J. (2013). Review of have we identified effective teachers? AND A composite estimator of effective teaching: Culminating findings from the Measures of

Effective Teaching Project. Retrieved from http://nepc.colorado.edu/thinktank/review-MET-final-2013

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

Sanders, W. L., & Horn, S. (1994). The Tennessee Value-Added Assessment System (TVAAS): Mixed-model methodology in educational assessment. *Journal of Personnel Evaluation in Education, 8*(3), 299-311.

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

Sawchuk, S. (2013, February 5). *Teachers' ratings still high despite new measures.* Retrieved from www.edweek.org /ew/articles/2013/02/06/20evaluate_ep.h32.html

Schacter, J., & Thum, Y. M. (2005) TAPing into high quality teachers: preliminary results from the Teacher Advancement Program Comprehensive School Reform. *School Effectiveness and School Improvement 16*(327), 327-353.

Spillane, J. P., & Kenney, A. W. (2012). School administration in a changing education sector: The US experience. *Journal of Educational Administration, 50,* 541-561.

Stake, R. (2004). Qualitative case study. In N. K. Denzin & Y. S. Lincoln (Eds.), *The Sage handbook of qualitative research (3rd ed.).* (pp. 443–466). Thousand Oaks, CA: Sage.

Steinberg, M. P. & Donaldson, M. L. (2016). The new educational accountability: Understanding the landscape of teacher evaluation in the post-NCLB era. *Education Finance and Policy, 11*, 340–359.

Sun, M., Mutcheson, R. B., & Kim, J. (2016). Teachers' use of evaluation for instructional improvement and school supports for such use. In Grissom, J. A. & Youngs, P. (Eds.), *Improving teacher evaluation systems: Making the most of multiple measures*. (pp. 102-115). New York, NY: Teachers College Press.

Sun, M., Penuel, W. R., Frank, K. A., Gallagher, H. A., & Youngs, P. (2013). Shaping professional development to promote the diffusion of instructional expertise among teachers. *Educational Evaluation and Policy Analysis, 35*(3), 344-369.

Taylor, E. S., & Tyler, J. H. (2011). *The effect of evaluation on performance: Evidence from longitudinal student achievement data of mid-career teachers*(No. w16877). National Bureau of Economic Research.

Thorn, C., & Harris, D. N. (2013). The accidental revolution: Teacher accountability, value-added, and the shifting balance of power in the American school system. In Anagnostopoulos, D. Rutledge, S. R., & Jacobsen, R. (Eds.), *The infrastructure of accountability*. (pp. 57-74). Cambridge, MA: Harvard University Press.

Tschannen-Moran, M., Woolfolk Hoy, A., & Hoy, W. K. (1998). Teacher efficacy: Its meaning and measure. *Review of Educational Research, 68*(2), 202-248.

Tyack, D. B., & Cuban, L. (1995). *Tinkering toward utopia*. Cambridge, MA: Harvard University Press.

Umpstead, R. R., & Kirby, E. (2012). Reauthorization revisited: Framing the recommendations for the Elementary and Secondary Education Act's Reauthorization in light of No Child Left Behind's implementation challenges. *West's Education Law Reporter 276*(1).

U.S. Department of Education. (2009). *Race to the Top Program executive summary.* Retrieved from www2.ed.gov/programs/racetothetop/executive-summary.pdf

Vinovskis, M. (2009). *From a nation at risk to no child left behind: National education goals and the creation of federal education policy*. New York: Teachers College Press, Columbia University.

Wayne, A. J., & Youngs, P. (2003). Teacher characteristics and student achievement gains: A review. *Review of Educational Research, 73,* 89–122.

Weisberg, D., Sexton, S., Mulhern, J., Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness.* The New Teacher Project. http://widgeteffect.org/downloads/TheWidgetEffect.pdf 2009

Yin, R. K. (2009). *Doing case study research (4th Ed.),* Thousand Oaks, CA: Sage Publications.

Youngs, P. & Grissom, J. A. (2016). Multiple measures in teacher evaluation: Lessons learned and guidelines for practice. In Grissom, J. A. & Youngs, P. (Eds.), *Improving teacher evaluation systems: Making the most of multiple measures*. (pp. 169-184). New York, NY: Teachers College Press.

Youngs, P. & Whittaker, A. (2016). The role of edTPA in assessing content-specific instructional practices. In Grissom, J. A. & Youngs, P. (Eds.), *Improving teacher evaluation systems: Making the most of multiple measures*. (pp. 37-50). New York, NY: Teachers College Press.

Yurkofsky, M. (2016). *Unpacking the panacea: Exploring how field-level pressures and inhabited actors influence the course of competition.* Paper session presented at the meeting of the American Education Research Association. Washington, D.C.