The Impacts of the Annual Professional Performance Review in New York State: Science Teachers’ and Administrators’ Perspectives

A Dissertation Presented

by

Jessica A. Mintz

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Science Education

Stony Brook University

August 2017
Stony Brook University
The Graduate School

Jessica A. Mintz

We, the dissertation committee for the above candidate for the Doctor of Philosophy degree, hereby recommend acceptance of this dissertation.

Angela M. Kelly, Ph.D.
Associate Professor, Department of Physics & Astronomy

Keith Sheppard, Ed.D.
Associate Professor, Department of Biochemistry & Cell Biology

Kenneth Lindblom, Ph.D.
Associate Professor, Department of English
Dean, School of Professional Development

Craig Markson, Ed.D.
Program Director, Educational Leadership
Assistant Dean, School of Professional Development

Charles Taber
Dean of the Graduate School
The Impacts of the Annual Professional Performance Review in New York State:

Science Teachers’ and Administrators’ Perspectives

by

Jessica A. Mintz

Doctor of Philosophy

in

Science Education

Stony Brook University

2017

Abstract

The goal of this study was to investigate New York State’s Annual Professional Performance Review (APPR) from the perspectives of secondary science teachers and their administrators. Examining their perceptions through interviews was insightful due to the subjects’ proximity to the teaching and learning processes. Five science teacher/administrator pairs from selected school districts were interviewed; all had varied ranges of experience and content certifications. The study also investigated the unintended consequences the teachers and administrators experienced using the APPR system.

This phenomenological research study lays the groundwork for making policy recommendations for science teacher evaluations. The goal was to examine teacher and administrator perceptions, the clarity and practicality of teacher evaluation reforms, as well as how motivational theory might incentivize teacher change through future reform efforts. Provisional coding was used in this study based upon prior research. The list of codes was generated using motivational theories applied to the design of teacher evaluation policy and reform implementation constructs.

Although the science teachers agreed with the importance of being evaluated, they generally viewed aspects of the process of quantifying their effectiveness as unclear, unfair, and flawed. The science teachers indicated that student variations in ability and performance were not considered when APPR was established. The science teachers recommended that the focus of teacher evaluations should be on content specific professional development. They proposed the establishment of peer review systems, teacher collaboration networks, and self-reflection documentation as means to improve their science teaching practices.

The administrators agreed that accountability was important, however, holding individual teachers accountable for student outcomes was not reliably measured through the APPR process. They recommended other forms of evaluative measures that would focus on professional development instead of an evaluative effectiveness score. Their recommendations involved
creating more time for science administrators to be teacher leaders rather than evaluators. The administrators proposed three main recommendations: 1) decreasing the number of formal observations and replacing them with frequent informal classroom visits; 2) peer-peer observations utilizing instructional rounds; and 3) educator involvement in the creation of improved science teacher evaluation, with implicit trust in the administrators to exert local control.
Dedication

To my family, both those who are with us and those who have passed, for your loving nature and support have encouraged my persistence and dedication to life’s endeavors.
**Table of Contents**

Chapter 1: Introduction ............................................................................................................. 1

Race to the Top Legislation ...................................................................................................... 2
New York Educational Law Revisions ..................................................................................... 2
Research Questions .................................................................................................................. 5
  1. How do secondary science teachers view the APPR evaluation system in terms of clarity,
     practicality, and motivation? ................................................................................................. 5
  2. How do secondary science administrators’ view the APPR evaluation system in terms of clarity,
     practicality, and motivation? ................................................................................................. 5
  3. What are the impacts of the APPR system on teachers and administrators? ...................... 5

Chapter 2: Review of Literature ............................................................................................... 6

Introduction ............................................................................................................................... 6
Motivation ................................................................................................................................ 7
Challenges of Motivational Incentives ................................................................................... 8
Theoretical Framework ............................................................................................................ 9
Teachers’ and Administrators’ Views of Teacher Evaluations ................................................. 11
Race to the Top ........................................................................................................................ 11
Measuring Teacher Quality .................................................................................................... 14
Holding Teachers Accountable for Student Performance ....................................................... 18
Alternative Practices in Teacher Evaluation ........................................................................... 20
Summary .................................................................................................................................. 21

Chapter 3: Methods .................................................................................................................. 22

Introduction ............................................................................................................................... 22
Research Questions .................................................................................................................. 22
Study Context ........................................................................................................................... 23
New York State Science Regents ............................................................................................ 24
Annual Professional Performance Reviews in New York State .............................................. 25
Research Design ..................................................................................................................... 29
Recruitment and Data Collection Procedures ....................................................................... 30
Pilot and Protocol Formation ................................................................................................... 31
Subjects and Context .............................................................................................................. 31
Contextual Descriptions of School Districts ......................................................................... 36
Development of Coding Scheme ............................................................................................. 37
Qualitative Data Analysis ........................................................................................................ 39
Validity and Reliability of Qualitative Data ........................................................................... 40
Summary .................................................................................................................................. 42

Chapter 4: Results and Findings ............................................................................................. 43

Introduction ............................................................................................................................... 43
Results ........................................................................................................................................ 44
Research Question 1 ............................................................................................................... 44
Research Question 2 ............................................................................................................... 53
Research Question 3 ............................................................................................................... 58
Within District Perspectives .................................................................................................... 66
Notable Findings ..................................................................................................................... 68
Summary .................................................................................................................................. 74

Chapter 5: Discussion and Implications .................................................................................. 75

Introduction ............................................................................................................................... 75
Table of Figures

Figure 1. Theoretical framework ........................................................................................................ 10
Figure 2. HEDI district “A” conversion chart ....................................................................................... 27
Figure 3. HEDI district “B” conversion chart ....................................................................................... 27
List of Tables

Table 1  *Comparison of SLO Scores for Similar New York School Districts* ........................................ 4
Table 2  *Summary of Point Criteria for Race to the Top* ........................................................................ 13
Table 3  *Select Suffolk County School District Demographics, 2011-12* ............................................ 24
Table 4  *Four APPR Categories based on Composite Score* ................................................................. 25
Table 5  *Teacher Evaluation Rating Categories* .................................................................................... 26
Table 6  *Example from Danielson’s Framework* ................................................................................... 28
Table 7  *Components of NYSUT Teacher Practice Rubric* ................................................................. 29
Table 8  *Teacher Descriptions and Respective School Characteristics* ............................................. 33
Table 9  *Administrator Descriptions and Respective School Characteristics* ................................. 35
Table 10  *Coding Scheme* ..................................................................................................................... 39
Table 11  *Science Teacher and Administrator Clarity and Practicality Perceptions* ...................... 69
Table 12  *Science Teacher and Administrator Response Summary* ................................................ 70
Table 13  *Summary of Recommendations* .......................................................................................... 84
Acknowledgments

The writing of a thesis has a way of dominating the life of not only the writer but also those individuals who share a home with that writer. It seems appropriate, at this point, to thank my husband, Robert. This long journey was started well before some major life changing events, our engagement, our marriage, and the birth of our amazing son, Jack. I want you to know that the emotional ties I have to all of those milestones individually outweigh the completion of this dissertation, for without these events, your support, and unconditional love, this project would have been destined to fail. As we have joked, now I have the time to start paying attention to you! As without Rob’s support to complete this milestone, it seems fitting that together we would like to thank the following people.

I would like to thank my advisor, Dr. Angela Kelly, whose meticulous attention to detail finally forced me to learn how to cite in APA format. She should be rewarded for her dedication as an advisor, for without her guidance, knowledge, and painstaking attention to detail this project would not be as it is today. Thinking back to the time when we had to choose an advisor, I am glad I choose you and that you accepted me.

I would like to thank the dissertation committee members, Dr. Keith Sheppard, Dr. Kenneth Lindblom, and Dr. Craig Markson. Dr. Sheppard experienced my persistence first hand by accepting me into the program after my second application and I thank you for the opportunity. Presenting and defending a dissertation is an anxiety-ridden event. The committee members eased that anxiety by keeping the defense calm and light while asking tough, thoughtful questions which was truly appreciated.

I would like to acknowledge the members of our research group: Jennifer Gatz, Caren Gough, Rich Gearns, Catherine Hantz, Kimberly Heal, Stephanie O’Brien, Linda Padwa, and Hope Sasway. We started this journey together and now we are all reaching the end. I would like to thank them for their support, critiques, powerful discussions, frustrating conversations, and ears for venting. You all are a wonderful group of people and I look forward to lasting friendships.

During this journey I was lucky enough to meet another group of fantastic educators who I now call friends. Shout out to the Long Island Association of Chemistry Teachers: Kathleen Dinota, Kristen Drury, Jason Gutlazier, Bryan Horan, Stephanie O’Brien, Martin Palermo, and Michael Shanzer! You are all amazing and have changed my teaching career in ways in which I never thought possible. Thank you for all of support and friendship throughout this journey.

Finally I would like to acknowledge my family. Thank you to my parents for instilling in me the values of persistence and dedication. Thank you to my sisters, Marianne, Christine and Stephanie, for believing in me throughout this journey.
Chapter 1

Introduction

Teacher evaluation and accompanying educational reforms have received considerable attention across the U.S. in recent years. National policies have revamped local teacher evaluation systems to include multiple measures of teacher performance while also providing pathways for professional improvement. In May of 2011, the New York State Education Department (NYSED) initiated a plan to use the New York State Regents assessment results as a factor in measuring the effectiveness of science teachers (New York State Legislature, 2010). The decision to use student assessment scores gathered from state generated exams, in part to measure a science teacher’s effectiveness, has been influenced by funds made available by Race to the Top, a program that offered competitive grant money to states that linked student achievement data to individual teachers (American Recovery and Reinvestment Act [ARRA], 2009).

New York’s Annual Professional Performance Review system (APPR) placed more emphasis on evaluating science teachers’ abilities with high stakes attached. Teachers may have been placed on a teacher improvement plan or terminated if found to be developing or ineffective, rather than using scores to provide meaningful professional development (Pallas, 2012). Research conducted by Cohen and Spillane (1992) revealed that efforts to improve instruction often resulted in insufficient changes. For reform efforts to be successful and sustainable, teacher acceptance and union support are vital (Hanaway & Rotherham, 2008). In order to improve the effectiveness of new policies, teachers’ experiences and perceptions should be an integral part of reform efforts (Datnow & Castellano, 2000).

Examining teacher and administrator insights regarding teacher evaluation is essential as teachers and administrators are the key players in the learning process (Goe, Bell, & Little, 2008). The fundamental goal of teacher evaluation is to improve instruction and student achievement (Namaghi, 2010; Taylor & Tyler, 2012). Since teacher quality has been linked directly to student progress (Stronge, Ward, Tucker, & Hindman, 2007), and teachers are the key players in the success of schools, their perceptions of the evaluation process will impact teaching practice and thus student learning. This qualitative study explores connections among teacher and administrator perceptions, reform implementation factors, and consequences of teacher evaluation systems.

New York changed the teacher and principal evaluation laws in March, 2012 (University of the State of New York, State Education Department, 2012). School districts across the state were required to submit their district plans before the start of the 2012-2013 school year. The numerous districts and the unique make up of the study, a suburban area of New York State, resulted in many different APPR plans within close proximity to each other. Some districts organized their APPR plans to assign teachers higher evaluation scores, while others made it rare for teachers to be considered ineffective (Forman & Markson, 2015). Science teacher
perspectives offered insights as to whether or not APPR had achieved its goal of providing educators with a valid score that could be used to promote professional learning. The underlying question that this study aimed to answer was whether teachers and administrators felt they were gathering information that would enhance teaching and learning.

This introductory chapter provides the background on the origination of New York’s APPR plans and the impacts Race to the Top had on policy changes associated with APPR. The chapter concludes with the introduction of the research questions. Chapter 2 presents a review of the literature pertaining to motivation and the design of teacher evaluations, challenges associated with motivational incentives, the theoretical framework of the study, teacher and administrator views of teacher evaluation, measuring teacher quality, teacher accountability, and alternative practices in teacher evaluation.

Race to the Top Legislation

The No Child Left Behind Act (NCLB) was the first education law to connect federal funding to testing and accountability (U.S. Department of Education, 2001). As a result of NCLB and the corresponding mandated annual student testing, schools were required to bring all students to higher standards by 2014 and “adequate yearly progress” (AYP) became the measure to determine whether schools were succeeding. Recently, President Obama’s Race to the Top called for national standards and student achievement data to hold schools and teachers accountable for receiving federal funds (U.S. Department of Education, 2009).

The Race to the Top competitive grant program is a part of the American Recovery and Reinvestment Act of 2009 (ARRA, 2009). President Obama signed the act into law as a way to stimulate the economy during a time of economic recession. The Race to the Top Fund was seen as an important investment in creating opportunities for the nation’s students in order to promote innovative educational reform (U.S. Department of Education, 2009). The Obama administration’s implementation of Race to the Top was designed for states to compete for a piece of the $4.3 billion initiative to transform education at the state level.

New York Educational Law Revisions

As a result of Race to the Top, many states have rewritten educational laws to become more competitive in the application process. For example, New York State added laws to include student performance data as part of a teacher’s evaluation to facilitate the process of removing ineffective teachers (New York State Legislature, 2010). By law, a teacher’s evaluation includes 40% of her composite score based on student achievement (Meyer, 2011). However, there has not been empirical evidence for this criterion to warrant such changes in state laws (Smarik, 2010). Some questioned the state’s ability to sustain such reforms and wondered whether the incentives to participate in the program during times of economic recession were politically and financially driven (Smarik, 2010).

Currently, there is little research regarding how school districts weigh different measures in summative teacher evaluation scores. The tools used to guide the evaluation process have considerable variation across districts (Kennedy, 2010). The Widget Effect, a study commissioned by the New Teacher Project in 2009, showed that among 12 school districts that were considered to have direct teacher observation evaluation systems, 99% of the teachers
received the highest rating (Weisberg, Sexton, Mulhern, & Keeling, 2009). This suggests a failure to differentiate teacher effectiveness.

**New York State teacher evaluations.** On May 28, 2011, New York’s Senate and Assembly voted to require that 40% of teacher evaluations were based upon student achievement and provided $20.4 million for a new longitudinal data system (Meyer, 2011). The new law implied that *Race to the Top* initiatives would lead to better education, making this type of teacher evaluation more permanent (Meyer, 2011). Although the law’s intent was to strengthen teacher quality and student learning in K-12 education, there has been limited research to suggest the reforms mandated by *Race to the Top* fostered a better educational system.

In the subsequent system in New York State, four rating categories for teacher performance replaced the prior three-category system. Teachers and principals were classified as “highly effective, effective, developing, and ineffective,” also known as HEDI scales (New York State Education Department [NYSED], 2015b). Teachers were annually reviewed for performance based on a composite score of student growth, student achievement, and other measures such as teacher observations.

Science educators and teachers of mathematics and English language arts (ELA) were evaluated differently in the newly developed system, since students were not tested every year in the same science content. New York State implemented a value-added model for ELA and mathematics in grades 3 through 8 (McCaffrey, Lockwood, Koretz, & Hamilton, 2003). Teachers who did not teach ELA or mathematics in these grades were evaluated on Student Learning Objectives (SLOs). SLOs were academic goals set by an educator at the beginning of each school year. The SLOs contained information on the learning content, the assessments used to measure the goal, the baseline level of the students in the class, the expected target by the end of the course, district-based HEDI ratings, and the rationale for why a teacher chose such targets. This implies that teachers not teaching subjects that used state generated value added growth scores had more freedom when setting up growth measures. The local school districts were in control of the growth and achievement targets, not NYSED as was the case for Regents science teachers.

The NYSED (2015b) outlined a structured plan for APPRs while the details were left to local districts. As a result of local control, teachers often received different grades if they were teaching the same content with the same types of students in a similar school setting but in a different district. For example (see Table 1), a teacher with 55% of their students meeting the target goal would be considered *effective* in both Bay Shore and Central Islip School Districts but would be considered *developing* in Brentwood School District (NYSED, 2013a).
The purpose of creating new evaluation systems was to recognize great teachers, to improve the practice of mediocre teachers, develop pedagogical skills, and to eliminate incompetent teachers. The reliability of the evaluation system may be questionable when the definition of effective teaching is different across school district borders. *Race to the Top* focused on four major criteria for educational reform, and one such criterion was to reward effective teachers. Teachers were supposed to receive a professional improvement plan or lose their jobs if they were found to be developing or ineffective, however, they would receive no reward if they scored highly effective or effective. Distinguishing great teachers from poor teachers is a complex undertaking and the research in this field is still inconclusive.

After this study began, New York changed part of the 2012 evaluation law. *Education Law 3012-d* was signed in April of 2015, which changed the evaluation system of classroom teachers and principals. The composite scores for teachers and principals were then based on two categories instead of the previous three-category system: 1) the student performance category, and 2) the observation category (NYSED, 2015b). Although the observation process did not change, the weighting of this category decreased. The previous law weighted the observation 60% of the composite score, whereas the newly implemented observation category was rated 50%. The category for student growth was increased to 50% rather than the original 40% for student growth and achievement. The four-rating categories to differentiate teacher effectiveness (highly effective, effective, developing, and ineffective) did not change and high stakes decisions for employment, supplemental compensation, tenure, termination, and promotion were still attached.

At the time of this study, the student performance category for science teachers was comprised of one mandatory subcomponent and an optional subcomponent. Student learning objectives were mandatory for science teachers of Regents courses, and the Regents exams had to be the underlying assessment. Districts could choose to add an optional subcomponent of

<table>
<thead>
<tr>
<th>School District</th>
<th>Percentages of Students Meeting Targets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bay Shore School District</td>
<td></td>
</tr>
<tr>
<td>Highly Effective</td>
<td>85-100</td>
</tr>
<tr>
<td>Effective</td>
<td>51-84</td>
</tr>
<tr>
<td>Developing</td>
<td>27-47</td>
</tr>
<tr>
<td>Ineffective</td>
<td>Below 47</td>
</tr>
<tr>
<td>Brentwood School District</td>
<td></td>
</tr>
<tr>
<td>Highly Effective</td>
<td>85-100</td>
</tr>
<tr>
<td>Effective</td>
<td>61 – 84</td>
</tr>
<tr>
<td>Developing</td>
<td>51-61</td>
</tr>
<tr>
<td>Ineffective</td>
<td>Below 51</td>
</tr>
<tr>
<td>Central Islip School District</td>
<td></td>
</tr>
<tr>
<td>Highly Effective</td>
<td>80-100</td>
</tr>
<tr>
<td>Effective</td>
<td>40-79</td>
</tr>
<tr>
<td>Developing</td>
<td>16-39</td>
</tr>
<tr>
<td>Ineffective</td>
<td>Below 16</td>
</tr>
</tbody>
</table>

Table 1
*Comparison SLO Scores for Similar New York School Districts*
district or school wide growth results using available state-provided growth scores that were locally calculated. For districts that did not choose the optional subcomponent, the mandatory subcomponent was weighted 100%, and for districts that chose the optional subcomponent, the mandatory subcomponent was weighted a minimum of 50% and the optional subcomponent no more than 50% (NYSED, 2015b).

**Research Questions**

The purpose of this study is to examine perceptions and potential consequences of the New York State APPR system for science teachers and administrators. This study intends to impact teacher evaluation reform in New York and nationally. The research questions are as follows:

1. How do secondary science teachers view the APPR evaluation system in terms of clarity, practicality, and motivation?

2. How do secondary science administrators’ view the APPR evaluation system in terms of clarity, practicality, and motivation?

3. What are the impacts of the APPR system on teachers and administrators?
   a. What challenges have been identified?
   b. What are the recommendations of science teachers and administrators on how to evaluate science teachers?
Chapter 2

Review of Literature

The introductory chapter provided the information about the origins of APPR from Race to the Top and why teacher evaluation systems in general have presented some challenges for teachers and administrators. New York State was announced as a winner of the Race to the Top competition in September, 2010 (U.S. Department of Education, 2010a). In March, 2012, New York changed the teacher and principal evaluation laws and school districts were required to submit APPR plans before the start of the 2012-2013 school year. The APPR plans were structured in nature but categorically relaxed resulting in many different district plans for teacher evaluation. This structure is further discussed in Chapter 3. The review of literature pertaining to science teacher and administrator perceptions of teacher evaluations is limited. Within this chapter a review of literature related to teacher evaluations is discussed.

Introduction

A recent wave of educational reforms has influenced changes in the way teachers are evaluated. As a result of Race to the Top, New York has shifted the focus of the evaluation to the individual educator rather than the school as a whole. Research on educational policy has called for the need to evaluate the practices and perceptions of teachers in order to understand the successes and challenges of policy changes (Datnow & Castellano, 2000). The most promising policy initiatives have consequences that depend on stakeholders’ interpretations and actions (McLaughlin, 1987). Researchers and policy makers must be attentive to teachers’ responses to change (Gitlin & Margonis, 1995), because successful implementation of new policy depends largely upon teacher acceptance (Hannaway & Rotherham, 2008).

The following review of literature is segmented into six major components. The first section discusses the theoretical framework of how teacher motivation and policy implementation factors are related to teacher acceptance of changes to educational policy. The second section focuses on teacher and administrator perceptions of teacher evaluations. The third section describes Race to the Top and the criteria for which states were awarded funds. This section discusses the purpose of the legislation and how it compares to previous reforms dictated by No Child Left Behind (NCLB) (U.S. Department of Education, 2001) and the Elementary and Secondary Education Act of 1965 (ESEA, 1965). New York State law revisions as a result of Race to the Top are discussed in the fourth section. The fifth section focuses on the current transformation of science teacher evaluations, highlighting teacher accountability using value-added models (VAM) and student learning objectives (SLOs). The sixth section describes research on teacher accountability related to student assessment scores.
Motivation

Designing teacher evaluation policy draws upon the degree to which data are used to reward or discipline teachers and how these same data are used to influence teachers to improve instruction (Bell, 2012). Motivation is the reason or desire a person has for acting or behaving in a certain way. Two overarching theories of motivation have been used in the design of teacher evaluation systems: the economics-based theory and psychology-based theory. These theories were evidenced in the design of teacher evaluations in New York State (NYSED, 2015a).

Economics-based theories. Economics-based theories of motivation make the assumption that teachers will respond to extrinsic incentives. Extrinsically motivated individuals participate in activities for reasons such as receiving a reward. This theory of motivation is visible in the requirements of Race to the Top (U.S. Department of Education, 2009) and these requirements have trickled down to New York State evaluation policies. Extrinsic incentives would include financially rewarding “highly effective” teachers and reprimanding teachers who are “ineffective”.

Allocation of extrinsic incentives does not come without obstacles (Laffont & Marimont, 2009). Principal agent theory can be used to explain some challenges that arise when distributing extrinsic incentives (Firestone, 2014). The idea is that the principal, or the person responsible for evaluations, demands compliance from the agent - the teacher - without the ability to monitor the teacher’s work effectively. The incentives are extrinsic because they are conditioned upon a measured outcome. One of the goals of Race to the Top was to “reward” great teachers, however, extrinsic incentives were implied but never implemented in New York State (U.S. Department of Education, 2009).

Examples of extrinsic incentives in education are the single-salary schedule, career ladder, and pay for performance. Many teachers throughout the U.S. follow a single-salary schedule (Podgursky, 2002). Under this salary schedule, teachers are rewarded for credits earned, degrees held and years of experience (Springer, 2009). Another example of extrinsic incentive in education is the career ladder, where teachers receive extra pay for doing additional work, such as leading professional development or curriculum writing as contracted by the school district (Firestone, 1991). The newest form of extrinsic reward is performance based pay or merit pay for teachers who achieve measurable ends (Firestone, 2014). Measuring the teacher’s influence on student learning can be demonstrated by improved student achievement on standardized tests and/or measures of teacher practices, such as strong pedagogical content knowledge, student evaluations, and rigorous classroom observations (Bill and Melinda Gates Foundation, 2013).

Psychology-based theories. Psychology-based theories rely on intrinsic incentives. When individuals are intrinsically motivated, they engage in an activity because they are interested in and find enjoyment in doing the activity. Teacher evaluations can be used to motivate and influence professional development. Individuals who are intrinsically motivated typically have self-efficacy and autonomy. When it comes to teacher evaluation, motivation theories pertaining to expectancy-value models of behavior fit the design of the evaluation systems in New York (Eccles & Wigfield 2002). Expectancies refer to the beliefs about how one will do on different activities and values have to do with the incentives or reasons for doing the activity.
**Self-efficacy.** Within the framework of social cognitive theory, self-efficacy is defined as “belief in one’s capabilities to organize and execute the courses of action required to produce given attainments” (Bandura, 1997, p. 3). Skaalvik and Skaalvik (2007) conceptualized and measured teacher-self efficacy. They defined teacher-self efficacy more specifically as “individual teachers’ belief in their own abilities to plan, organize and carry out activities required to attain educational goals” (Skaalvik & Skaalvik, 2007, p. 612). Self-efficacy has been associated with teachers’ classroom behaviors and performance (Rimm-Kaufmann & Saywer, 2004).

When teachers believe they can implement a lesson resulting in the intended student outcomes, their motivation often becomes stronger (Tschanen-Moran, Woolfolk Hoy, & Hoy, 1998; Vroom, 1964). Teacher beliefs and attitudes have been closely associated with their classroom behavior and educational classroom practice (Richardson, 1996), and these beliefs provided the framework for which teachers make decisions daily within their classrooms (Calderhead, 1996). For example, if a teacher encounters the same misunderstanding of a particular science concept year after year, he is able to address this misunderstanding during the lesson. The teacher is comfortable addressing the misunderstanding and will have high confidence that the students will leave with a more scientific conceptualization. If the teacher were unfamiliar with common content specific misunderstandings or did not believe she needed to address misconceptions in her instruction, she may feel inadequate as an educator when students perform poorly on the content exam or blame students for lack of studying.

The relationship between teacher self-efficacy and student performance is insightful. Teachers who have strong self-efficacy have been more likely to improve the self-efficacy of their students (Midgley, Feldaufer, & Eccles, 1989), and teachers have felt more efficacious when their students performed well (Ross, 1992). This relationship is in delicate balance and requires further study. Teacher evaluations that carry extrinsic motivational incentives might disrupt this balance because teacher self-efficacy may be altered when extrinsic motivators outweigh the intrinsic incentives.

**Autonomy.** Teachers who have autonomy have often internalized the goals of administrators, principles or educational leaders if they found those goals reasonable. These teachers required no additional incentives to motivate them (Deci & Ryan, 1996). Locke and Latham (2002) found that when the goals came from a valid authority figure, such as an esteemed principal or administrator, specific and challenging goals produced greater effort. Teachers have had great autonomy throughout history (Corwin & Borman, 1988), and with the rise of accountability polices, teachers still reported that their work was less influenced by these policies and influenced more by student and peer feedback (Firestone, Nordin, Scherbakov, Blitz, & Kirova, 2014).

**Challenges of Motivational Incentives**

Current research suggests that combining incentives into one teacher assessment policy is difficult. The Teacher Advancement Program (TAP) makes these challenges evident. TAP is a national educator effectiveness model that includes opportunities for career advancement, professional growth, and accountability compensation. TAP considers three elements for teacher educational success: multiple career paths, instructionally focused accountability, and
performance-based pay. TAP mixes extrinsic and intrinsic rewards, yet has resulted in only small improvements in student achievement and teacher retention (Glazerman & Seifullah, 2012; Springer, Ballou, & Peng, 2008). Other researchers have reported the interaction impacts of extrinsic and intrinsic incentives have been mixed (Ryan & Deci, 2006).

Firestone (2014) suggested three main challenges when combining incentives: 1) intrinsic incentives are challenged by extrinsic incentives, 2) high-stakes exams may not provide teachers with productive feedback, and 3) the time required for administrators to collect evidence of effectiveness competes with the time to promote teacher self-efficacy. First, autonomy required to support intrinsic incentives has been undermined when the rewards were predictable, as in pay for performance or merit pay (Lepper & Henderlong, 2000). Research has shown that extrinsic rewards that provided information about one’s competence enhanced intrinsic motivation when that information was valuable, less tangible, and unexpected (Lepper & Henderlong, 2000).

Secondly, high stakes exams may not be the ideal tool to provide teachers with feedback that will foster their sense of competence. Weiss (2012) has reported the difficulty in designing assessments that could monitor the distribution of extrinsic rewards while simultaneously creating intrinsic ones. Teachers used accountability assessment data productively when the data did not have high stakes attached (Jennings, 2012). In those cases, the information gathered would not be used to reward or punish the educator. However, when the data were obtained quickly and teachers gained an understanding of the content learning challenges faced by their students, the data were used more effectively (Jennings, 2012).

Lastly, the time required for administrators to collect the necessary information to distribute extrinsic rewards has competed with time needed to create the working conditions that would increase teacher efficacy (Firestone, 2014). Research has shown that administrators have less time to provide support to teachers in need (i.e., untenured), provided less productive observation feedback, and distributed responsibilities to others to minimize the increased demands of accountability policies (Curtis, 2012; Gonzalez & Firestone, 2013; Milanowski & Kimball, 2003).

Theoretical Framework

Similar to the studies of Tuytens and Devos (2009) and Jiang, Sporte, and Luppescu (2015), this study employed theoretical constructs from Fullan (2001) as a way to delineate teacher perceptions of evaluations. Fullan (2001) identified two significant reform implementation factors that relate to how teachers change their practices and beliefs in response to new policies. These factors are clarity and practicality. Clarity addresses how explicitly policy goals are operationalized, and practicality refers to whether teachers consider the new policy feasible to implement (Fullan, 2001).

Doyle and Ponder (1977) defined three constructs of practicality in teacher decision making: instrumentality, congruence, and cost. The construct of instrumentality refers to how the policy provides teachers with the means and the process for change. Congruence refers to the policy being consistent with the educator’s existing teacher practices and beliefs. Finally, cost denotes the time, resources and the knowledge it takes to implement the change. This study used these factors as an exploratory lens for semi-structured interview questions that elicited New York State teachers’ and administrators’ perceptions of the APPR evaluation methods (see interview protocols in Appendices A and B, respectively).
A fundamental question in teacher evaluation reform is how much weight should be given to teacher evaluation data to reward or penalize teachers, and how these evaluations are used to improve instruction (Bell, 2012). The two parts to this question are informed by different theories of motivation, as summarized in Figure 1. The first part relies on extrinsic incentives to motivate teachers to improve instruction. The second part of this fundamental question focuses on intrinsic incentives, whereby teachers self-reflect and engage in professional development to improve instruction (Firestone, 2014). Research has shown that it is difficult to reconcile these two theories since extrinsic motivation incentives weaken intrinsic incentives (Ryan & Deci, 2006). This study draws upon these two motivation theories to investigate stakeholders’ perceptions of evaluation reform in specific cases in suburban New York State.

Figure 1. Theoretical framework (adapted from Firestone, 2014; Fullan, 2001).

It is hypothesized that positive teacher perceptions will correspond to successful implementation of teacher evaluation reform (Figure 1). Positive teacher perceptions of evaluation reform efforts will be connected with more heavily weighted intrinsic incentives over extrinsic ones, and teacher acceptance of evaluation reform will be based on the degree to which they understand the policy and its practicality. Teacher “buy-in” refers to the teacher’s understanding of a new reform effort and the successful implementation of that effort from the bottom (the level of the teacher) to the top (the state policy level). A successful implementation would result in the change of teacher practices and beliefs to fulfill state policy goals. The “buy-in” would indicate the science teacher’s support of a new, meaningful system, not pressure to force teachers to support a flawed one.
Teachers’ and Administrators’ Views of Teacher Evaluations

Teacher evaluation has sparked debate over how to hold teachers accountable for student outcomes and the best practices to implement associated outcomes effectively (Baker et al., 2010; Papay, 2012). Some research has shown teacher evaluation reform efforts offered promise for improving student learning (Taylor & Tyler, 2012), however, little research has been done on teacher and administrator views and understandings of these reforms. Examining teacher and administrator perceptions regarding teacher evaluations is critical since teachers and administrators are closest to the learning process (Goe et al., 2008).

Teachers have responded in various ways to educational reforms. Some educators actively resisted change, some engaged with enthusiasm, and others complied but were not actively engaged in reform efforts (Datnow, 1998). Teachers often resisted educational policy change if they felt that reforms did not match their interests, beliefs or values (Muncey & McQuillian, 1996). These educational policy changes have often been met with intense and negative emotional teacher reactions (Schmidt & Datnow, 2005). Spillane (1998) found that teachers’ perceptions of new policies were influenced by structural and social conditions of the school and relationships with school administration.

Educators generally believe that evaluation is not a useless practice (Clipa, 2015). Teachers have expressed that evaluations should be used to measure and develop their pedagogical skills (Marzano, 2012). Zimmerman and Deckhert-Pelton (2003) found, based on survey responses by teachers in Florida, relationships among evaluators’ communication skills, teachers’ commitment to professional development, their pedagogical knowledge and pedagogical content knowledge, and teachers’ perceptions that the process was an effective evaluation tool. Another study indicated that teachers did not have confidence in the evaluation system to improve their craft because they had little collaboration with their evaluators (Wacha, 2013). Tuytens and Devos (2009) found positive perceptions of new evaluation policy in Belgium yet concerns regarding how the school implemented the change. A recent study on teacher perceptions of evaluation reform in Chicago showed that teachers were concerned about the addition of student growth as a part of their evaluation but found the observation process provided useful feedback (Jiang et al., 2015).

As a result of Race to the Top, many principals and administrators assumed the responsibility of implementing new teacher evaluations. Principals voiced concerns about the perceived inequities of the teacher evaluation system and its impact on teachers, describing the evaluation process as inconsistent and subjective. These concerns were intensified when their knowledge and time management issues were not addressed during the transition to the new evaluation policy (Derrington & Campbell, 2015). The following section will describe recent federal legislation that resulted in the new evaluation system in New York State.

Race to the Top

The purpose of Race to the Top. President Obama initiated the Race to the Top competitive grant program as a part of the American Recovery and Reinvestment Act of 2009 (ARRA, 2009). The grant program had four major goals: 1) to ensure that every classroom had great teachers and schools were lead by great administrators, 2) provide families with information to improve their local schools and information for teachers to improve student learning, 3) implement college and
career ready programs, and 4) improve student learning and achievement in underperforming districts (U.S. Department of Education, 2009).

To some, the manner in which states were encouraged to participate in Race to the Top was controversial. States lost Title I benefits, previously outlined in the Elementary and Secondary Education Act (ESEA) of 1965 and reauthorized in NCLB of 2001 (U.S. Department of Education, 2001), if they chose not to participate in the new Race to the Top competitive grant program and adopt new national standards (Barnes, 2011). Title I is a federal aid program that provides financial assistance to schools that have a high percentage of students from low-income families (ESEA, 1965). According to the National Center of Education Statistics, 62,305 schools out of 98,706 were eligible for participation in the $14.5 billion Title I program in the 2008-2009 school year (U.S. Department of Education, 2010b). Race to the Top provided an additional $10 billion in new funding to allow schools with underprivileged populations to close the achievement gap (U.S. Department of Education, 2009). Largely populated states with big cities dependent on Title I funding, e.g. New York and Chicago, had to enter the competition to avoid the risk of losing significant revenue. New York City schools received approximately $600 million from the Title I program as part of their $19.2 billion operating budget in 2012 and the loss of these funds would be significant (New York City Department of Education, 2013).

State criteria for Race to the Top grant funding. Winning states were awarded funds in three phases: Phase 1, for states that were ready to apply for the January 2010 deadline; Phase 2, for states that needed more time to apply with an extended deadline of June 2010; and for Phase 3, applications were due by November 2011 (U.S. Department of Education, 2009). Only two states, Tennessee and Delaware, were awarded funds in Phase 1. The District of Columbia, Florida, Georgia, Hawaii, Maryland, Massachusetts, New York, North Carolina, Ohio, and Rhode Island followed in Phase 2 (The White House, Office of the Press Secretary, 2010). Phase 3 funding was awarded to Arizona, Colorado, Illinois, Kentucky, Louisiana, New Jersey and Pennsylvania (The White House, Office of the Press Secretary, 2011). Forty-six states in total have applied for Race to the Top funds, and to date, only 19 states have been awarded federal assistance. Multiple states have sent in applications more than once, including Alabama, Arkansas, Connecticut, Iowa, Michigan, Missouri, Nebraska, New Hampshire, New Mexico, South Carolina, Oklahoma, Utah and Wisconsin, without securing Race to the Top funding.

The Race to the Top fund focused on four major criteria for educational reform. For states to be awarded funds they must have had plans in place to meet the core reform areas, as described below:

- The state must adopt new standards and assessments to succeed in college and the workplace and to compete in the global economy; building data systems that measure student growth and success, and inform teachers and principals about how they can improve instruction; recruiting, developing, rewarding and retaining effective teachers and principals, especially where they are needed most; and turning around our lowest-achieving schools (U.S. Department of Education, 2009, p. 2).

State proposals were evaluated using a 500-point rubric by a panel of externally trained reviewers. The U.S. Department of Education sent out a request for peer reviewers to help make judgments for which applicants would receive grant money. More than 1500 people applied to be
peer reviewers and 58 “highly qualified” individuals were chosen. These individuals had no conflicts of interest and the group was comprised of retired teachers and principals, district superintendents, professors, education supporters, and business leaders (U.S. Department of Education, 2010b). The Department did not determine a cut off score or number of winners for each phase. However, looking through the winning state’s data, Phase 1 had two winners, with the Delaware coming in first with a score of 454.6 points and Tennessee scoring 444.2 points. Phase 2 had more winners with the lowest winning score at 440.8 points.

The number of points available for each of the criteria is summarized in Table 2. For applicants to be competitive, they had to include STEM components, with an emphasis on preparing students for advanced study, incorporating STEM across the disciplines, and cooperating with surrounding research facilities, museums, and universities. Each category was broken down into specific items and point values. For states to win points in the “State Success” criterion, they had to implement reforms in the four educational reform categories and describe plans to enforce and sustain their proposed ideas. Common standards and common high-quality assessments had to be adopted. Data systems had to be created that could track and report student achievement in a statewide longitudinal system for the purpose of instructional improvement. Educators were to be evaluated on performance and good teachers were to be evenly distributed throughout the state. The states must have had plans for better teacher preparation programs with alternative certification pathways and teacher support (U.S. Department of Education, 2010b).

<table>
<thead>
<tr>
<th>Criteria</th>
<th># of Points</th>
<th>% of Total Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>State Success</td>
<td>125</td>
<td>25%</td>
</tr>
<tr>
<td>Standards &amp; Assessments</td>
<td>70</td>
<td>14%</td>
</tr>
<tr>
<td>Data Systems</td>
<td>47</td>
<td>9%</td>
</tr>
<tr>
<td>Great Teachers &amp; Leaders</td>
<td>138</td>
<td>28%</td>
</tr>
<tr>
<td>Turning Around Low Performing Schools</td>
<td>50</td>
<td>10%</td>
</tr>
<tr>
<td>General</td>
<td>55</td>
<td>11%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Priorities</th>
<th>Absolute</th>
<th>3%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compressive Approach to Education</td>
<td>Absolute</td>
<td>3%</td>
</tr>
<tr>
<td>STEM</td>
<td>Invitational</td>
<td></td>
</tr>
<tr>
<td>Innovations for Early Learning</td>
<td>Invitational</td>
<td></td>
</tr>
<tr>
<td>State Longitudinal Data Systems</td>
<td>Invitational</td>
<td></td>
</tr>
<tr>
<td>P-20 Coordination</td>
<td>Invitational</td>
<td></td>
</tr>
<tr>
<td>School Level Reform Conditions</td>
<td>Invitational</td>
<td></td>
</tr>
</tbody>
</table>

**Issues with adoption of Annual Professional Performance Reviews.** One critical issue with APPR is that the assessment of teachers and the development of their pedagogical skills are two different goals (Marzano, 2012). Evaluations designed to measure a teacher’s ability differ from those used primarily for teacher development. Marzano (2012) found the majority of educators believed teacher evaluation systems should include both measurement and development but that development should be the more important purpose. APPR in New York placed more emphasis on grading teachers’ ability with high stakes testing rather than using the grade to develop more effective teachers (Pallas, 2012).
Measuring Teacher Quality

Teacher quality. Many advocates for accountability reform have made the assumption that higher performing schools prepare students for higher education whereas lower performing schools do not prepare students for college. Some studies have suggested that limited metrics used to hold schools accountable narrow the curriculum (Rothstein, 2009; Rothstein, Jacobsen & Wilder, 2009). The literature on this topic is limited since assessment scores were not commonly used to evaluate teacher effectiveness prior to Race to the Top. Some studies have urged educators to proceed with caution before putting a model of this nature into practice (Hinchey, 2010).

There is no agreed upon ideal process for evaluating teachers and very little research specifically dedicated to science teacher evaluations, even with the complexities of science content specializations. There is agreement, however, that no one measure is a reliable proxy of teacher effectiveness but rather many factors contribute to the overall skillset. The amount of teacher education, years of classroom experience, content knowledge, pedagogical knowledge, pedagogical content knowledge, classroom interactions, classroom activities, and teacher involvement in the community can all be linked to teacher effectiveness, however, their importance has been minimized by certain evaluation methods (Hinchey, 2010).

There has been some argument about the purpose of teacher evaluations. Is the purpose to measure teacher quality or develop teachers? Marzano (2012) summarized three characteristics that focus teacher evaluations on professional development: 1) the system includes all elements that research has identified as associated with student achievement and corresponding classroom strategies; 2) the system includes a rubric to guide and track teacher development by setting and meeting teacher goals (Marzano, Frontier, & Livingston, 2011); and 3) the system should acknowledge and reward growth. These guidelines were considered in the present study by questioning teachers and administrators about the effectiveness of APPR.

Teacher evaluation. Teacher evaluation has been defined as “collecting and using information to judge” teacher quality (Darling-Hammond, Wise, & Pease, 1983, p. 290). The goal of teacher evaluations has been to improve teaching (Namaghi, 2010; Taylor & Tyler, 2012), and student learning has been most impacted by the effectiveness of the teacher over any other variable including class size (Gorden, Kane, & Staiger, 2006; Hanushek, Kain, O’Brien, & Rivkin, 2005). There has been considerable debate over how to evaluate the teacher’s effect on student achievement (Piro, Wiemers, & Shutt, 2011), including a lack of a clear description of the term “teacher quality” (Kennedy, 2008). There is a general consensus that no one measure is representative of teacher effectiveness because teaching and learning have been influenced by many factors (Hinchey, 2010).

The most common way to measure teacher effectiveness is through subjective observation or formative evaluation. A formative observation is meant to shape, form, or improve teachers. This type of evaluation has been criticized because it relies on one evaluator’s perception of teacher effectiveness without data to support his/her interpretation (Strong, Gargani, & Hacifazlioglu, 2011). Practice-based teacher evaluation is arguably the most direct evidence of a teacher’s ability to affect student learning and would suggest that observations be heavily weighted when calculating a teacher summative score. The Measure of Effective Teaching (MET) Project (Bill and Melinda Gates Foundation, 2011) has shown positive correlations among five different observation rubrics and student achievement. Reliability of scores was increased when multiple
observations were averaged and student surveys were added (Bill and Melinda Gates Foundation, 2011). Recent policy changes often focused on summative evaluation, aiming to use data to inform judgments of teachers rather than their professional development.

States have developed and implemented new teacher evaluations systems because methods used in the past have been unable to measure teacher quality due to an inability to distinguish effective from ineffective teachers (Bill and Melinda Gates Foundation, 2011; Toch & Rothman, 2008; U.S. Department of Education, 2009; Weisberg et al., 2009). Several studies have shown that the current practices of teacher compensation based on years of experience and academic credentials have had little impact on student learning (Goe & Stickler, 2008; Hanushek et al., 2005; Walsh & Tracy, 2004). Further research has shown that teacher evaluations ranked most teachers as satisfactory or good, but failed to recognize great teachers and offered little professional development for poor performing teachers (Weisberg et al. 2009).

Teacher evaluation reform. The U.S. has attempted to establish a system that identifies struggling schools and rewards those that showed innovative and effective practices. Such a system would support educators in improving their performance, as well as increasing student achievement while preparing them for the workforce and post-secondary study. As a result of NCLB, high stakes testing has been a key feature of state and federal accountability systems (U.S. Department of Education, 2001). Since NCLB (U.S. Department of Education, 2001) and Race to the Top (U.S. Department of Education, 2009), the connection between teacher effectiveness and student achievement has prompted reform efforts in the way teachers are evaluated (Strong et al., 2011).

The reauthorization of the Elementary Secondary Education Act (ESEA) in 2002 set the foundation for increased accountability established by NCLB (U.S. Department of Education, 2001). The main goal was to improve student achievement. NCLB required schools to meet preset proficiency rates and show adequate yearly progress (AYP) for all students and student subgroups (Franciosi & Regaldo, 2004). Since NCLB legislation was enacted, schools were held accountable for student test scores, not individual teachers. Schools that were considered “poor performing” have lost funds and faculty, and risked being taken over by the state government (Herman, Webb, & Zuniga, 2007). In 2009, the Race to the Top initiative prompted states to expand educational accountability to include student achievement as metrics in the evaluations of teachers and administrators (U.S. Department of Education, 2009).

Value added measures. Common measures of teacher qualification, such as years of experience or advanced degrees, have provided little information about individual teacher effectiveness and value-added measures (VAMs) hold more promise for measuring teacher quality (Corcoran, 2010). Value-added models have added student growth to the teacher’s evaluation in an attempt to control for outside factors that affect student achievement (Harris, 2009). School report cards have been used to record student achievement at one point in time, on one particular exam, without knowing the student’s baseline performance. VAMs attempt to take into account the starting and finishing point of a student’s achievement while receiving instruction from a teacher in an individual subject.

VAMs use inferential statistical analyses to show teacher effectiveness through normalized gains in standardized test scores. Controversy lies in the fact that policy makers have used VAMs that depended on untestable statistical models and have made inaccurate inferences regarding
teacher effects and assumptions about measured student learning (Newton, Darling-Hammond, Haertel, & Thomas, 2010).

**Value added measures debate.** The debate on the use of VAMs is widely divergent and schools have been caught in the middle while having to utilize these measures in evaluations (DiCarlo, 2012). Opponents argued that VAMs are unreliable and invalid and should not be used in high-stakes evaluations that guide employment, while supporters suggested the measure needs more weight. Regardless of confidence on either side, there is no empirical evidence that using VAMs has improved teacher performance and student achievement (DiCarlo, 2012).

There are many concerns surrounding value-added assessment. The complicated statistical analysis makes understanding the process very difficult. Taylor and Nguyen (2006) argued that VAMs should not be used as a “performance indicator” but rather used to determine why different schools show different gains. Researchers have been concerned that VAMs depend on test scores and the nature of the constructs being measured might influence the estimated effects (McCaffrey et al., 2003).

Much literature has called for more verifiable evaluations of VAMs and the sensitivity of the measure of a teacher’s effect on students’ educational growth. Large variations in estimated effects with different mathematical achievement levels showed that teacher performance using VAMs was sensitive to the ways in which student achievement was assessed. Because of these variations, the interpretation of estimated teacher effects and teacher performance might rely solely on skills measured by the test (Lockwood et al., 2007).

Koedel and Betts (2007) showed that individual teacher VAMs were unstable over time. McCaffrey, Sass, Lockwood, and Mihaly (2009) found that reliability coefficients increased by 40-60% when aggregating data across two years and another 18-23% when including a third year. Since reliability increased when using multiple years of data, new teachers could not receive fair value-added estimates. The study also found that when student and school effects were omitted from value added formulas, the stability improved. Hanushek and Rivkin (2010) agreed that in order for teacher evaluations to be fair, they must include other observable forms of evaluation, such as peer or administrator observations. This would prevent penalization of teachers in more challenging classrooms.

There have been concerns that value-added measures may be biased. Teachers have been assigned students with varying characteristics, such as level of parental engagement, which could affect the student’s performance (Rothstein, 2009). No single teacher accounts for a student’s success and achievement. Teachers from the past play a role in the future abilities of a student and several teachers can interact to influence learning. Some studies have shown VAMs are correlated with socioeconomic characteristics (Newton et al., 2010; Taylor & Nguyen, 2006). Teachers who might teach in a more affluent area may be considered more effective simply because the student has more resources and privileges.

Kane and Staiger (2012) concluded that teacher value-added scores were correlated with principal evaluations, however, it was not clear how principals were assigning teachers before the experiment. The MET Project also showed similar correlations (Bill and Melinda Gates Foundation, 2013). The findings of these studies revealed that higher student achievement gains were associated with higher observation scores. VAMs looked promising because they were replicated and correlated to principals’ observations of their teachers, but the measures were imprecise and did not seem to reflect teacher performance (Harris, 2009).
Recent studies of the New York, Texas, and Massachusetts educational systems found that parts of the curriculum were not tested on summative assessments (Holcombe, Jennings, & Koretz, 2010; Jennings & Bearak, 2010), and teachers were aware of systematic repetitions and exclusions that appeared on exams and adjusted to teaching to that format (Corcoran, 2010). Inconsistencies were common and countered the intended goals of VAMs. Corcoran (2010) compared value-added calculations for two standardized tests administered in Houston and found that teachers who scored in the top category of the Texas Assessment of Knowledge and Skills Test ranked in the lowest category on the Stanford Achievement Test.

The majority of valued-added systems have failed to distinguish teacher influences from school level effects on student achievement (Corcoran, 2010). Research has shown that students performed better when teachers were surrounded by a greater number of effective colleagues (Jackson & Bruegman, 2009), and principal leadership played a significant role in student outcomes and value-added scores (Clark, Martorell, & Rockoff, 2009).

VAMs that are in practice in New York have not included the majority of educators. New York State added VAMs to mathematics and English Language Arts (ELA) in grades four through eight (NYSED, 2015b). However, few educators teach subjects that assess student learning with a summative standardized test in mathematics and English. High school science teachers were not evaluated using a value added system. They were evaluated by setting student exam targets instead of state assigned growth score. State tests were often predictable in content and format allowing teachers to be rewarded for mastering exam idiosyncrasies (Corcoran, 2010).

The state of teacher evaluation reform. In a report by the Center of American Progress, early adopting states of new teacher evaluation systems were compared to provide an overview of teacher evaluation reform efforts (McGuinn, 2012). Teacher evaluation systems in Colorado, Delaware, New Jersey, Pennsylvania, Rhode Island and Tennessee were analyzed and assessed. McGuinn (2012) identified challenges and lessons from these six states to guide future reform efforts in other states.

The role of state educational agencies differs from state to state. The traditional focus of the state system is on compliance and accountability, which has made local districts cautious about being open with difficulties in implementing reform efforts. State education departments have been undergoing radical changes, shifting from compliance monitors to school-improvement organizations. This shift was extremely time consuming and states have reported struggling with the time constraints of the Race to the Top initiative. For most states, the tight fiscal climate and the time limits for federal and foundation grants posed problems for long standing implementation of new teacher evaluation systems. Lack of communication between local educational agencies, state agencies, and the U.S. Department of Education have also been reported (McGuinn, 2012).

Tennessee was one of the first winning states for the Race to the Top competitive grant program and has become a model for other states applying for federal funds (The White House, Office of the Press Secretary, 2010). The implementation and measurement of Tennessee’s reforms have been complicated and inconclusive. The Tennessee Value-Added Analysis System (TVAAS) and Tennessee Comprehensive Assessment Program (TCAP) used student data to show achievement and progress in the classroom as they related to teacher quality. However, there were some notable issues with these assessments.

Tennessee reported 75% of their teachers scored “above expectations,” which was higher than the predicted distribution, however, these scores aligned with measurements of student
growth and achievement. Teachers who had effectiveness scores of “below expectations” did not see this type of alignment (Tennessee Department of Education, 2012). These results showed that evaluation systems were not distinguishing between highly effective and effective teachers.

Administrators of the TVAAS reported their system had data to support the superior job their teachers had done and did not rely on data that marked students as proficient. In 2011, the average score in 8th grade mathematics (274) was lower than the national average score (283) and the state only saw a one-point increase from 2009. The same trend was seen in 8th grade reading. The Nation’s Report Card of Tennessee showed results in science growth as “not significantly different” between the years of 2009 and 2011 and only 64% of their students scored at the NAEP basic level with 31% at the proficient level (National Center for Education Statistics, 2011). This information suggested that the evaluation system had not yet increased student proficiency in mathematics and reading.

McGuinn (2012) discovered that the biggest challenge in the six state case studies was the fact that the majority of teachers did not teach in tested subject areas and as a consequence, standardized student achievement data were not available to be used in their ratings. These states created student learning objectives (SLOs) to solve this problem.

A recent study out of New York showed inverse relationships between teachers who were classified as “effective” and student achievement on the New York State English Language Exam and Mathematics Exam among student’s grades 3-8 (Forman & Markson, 2015). Forman and Markson (2015) discovered that the strongest correlation with students performing on or above grade level on these exams was the “ineffective” teacher-rating category and the authors implied that the “effective” rating had become the new “ineffective” rating in New York under the new evaluation law. The under reporting of “ineffective” teachers from this study may have been caused by districts eliminating underperforming educators before they could were eligible for tenure.

**Holding Teachers Accountable for Student Performance**

**Science assessments.** Because of the serious potential consequences for teachers, high-stakes assessments need to be evaluated to see if they are meeting their intended purposes. The phenomenon known as “teaching to the test” has created a “default” philosophy in education. Gunzenhauser argued, “default philosophy is one that places inordinate value on the scores achieved on high-stakes, rather than on the achievement that the scores are meant to represent” (Gunzenhauser, 2003, p. 51). Teachers may spend time “drilling” for test preparation or eliminating important concepts because they are not tested.

School achievement can be measured by using assessments that are based on the content being taught by the educators and through district made exams (Gunzenhouser, 2003). NCLB and Race to the Top have had inadvertent consequences consistent with the “default” philosophy of education. These policies placed significant importance on standardized tests since they had been used as an evaluative tool for the school and the individual teacher. The emphasis on test scores can have undesirable consequences that undermine teacher development, for example, systemic cheating and encouraging struggling students to drop courses (Nichols, Berliner, & Noddings, 2007). Assessments that should be useful in monitoring educational standards have often lost their consistency when high-stakes have been paired with them (Linn, 2000). In New York State, science exams created by the Board of Regents have been given high stakes value.

All of the sciences taken in New York at the high school level conclude with a high stakes summative exam known as the Regents exam. The living environment Regents exam is the only
science test required for graduation in New York and is often less difficult and more lenient on its grading curve. Chemistry and physics may seem difficult in comparison to living environment because the content is often considered more rigorous and most students need to study consistently to earn passing grades. Earning a raw score of 40 out of 85 possible points on the June 2012 living environment Regents exam resulted in a passing grade of 65% (NYSED, 2013b). Essentially, the student needed to know less than half the information to earn a passing score. However, on the chemistry Regents exam for the same year, the student needed to score a 49 out of 85 points to earn a passing grade of 65% (NYSED, 2013b). These results imply that living environment teachers are more effective than chemistry teachers while this is a questionable assumption.

Assessments can only focus on a specific objective, and multiple types of assessments must be used to make inferences about what students know (Pellegrino, 2012). The current national model of assessments assigns accountability to students, teachers and districts. According to Pellegrino (2012), this accountability was based on a large-scale assessment that could not possibly serve more than one purpose. Pellegrino (2012) suggested multiple coordinated assessments could be used to fully make valid inferences on what students know.

Major themes in science education include scientific inquiry, experimental design, and, more recently, applications in engineering. According to Pellegrino (2012) each of these themes serves different objectives and cannot be tested by a summative assessment like the Regents exam. Students should complete more than one test once a year in one sitting. Contemporary science exams still use the multiple-choice format; this structure has been criticized for its inability to evaluate the scientific process (Champagne & Newell, 1992), conceptual understanding (Tamir, 1998), reasoning ability (Frederiksen, 1984; Resnick & Resnick, 1992), and problem solving ability. These types of exams have encouraged teachers to emphasize unrelated facts during instruction (Tamir, 1998).

Day and Matthews (2008) analyzed living environment Regents exams from June 2004 through August 2006, looking for behavior classifications that supported inquiry components. The research showed that 21% of the questions assessed some sort of inquiry based knowledge. The Part D questions, which were intended to be the inquiry section of the exam, tested recall knowledge more often than inquiry. This representation of inquiry questions was much lower than what the rating guide recommended (Day & Matthews, 2008).

**Student learning objectives.** One of the challenges of practice-based science evaluations in high school is content specialization, since teachers in New York State may be certified in biology, chemistry, physics, or Earth science, and there are specific Regents exams for each subject. The first few states to implement data based targets have struggled with developing proxies for student achievement since the majority of teachers did not teach in tested subject areas (McGuinn, 2012). Many states developed SLOs that were not based on VAMs because of the difficulties of implementing value-added measurement at the high school level. High school science falls into this category. An SLO is a quantitative target of student growth that a teacher set for students (Tyler, 2011). The teacher first established a baseline for student knowledge at the beginning of the academic year. A district generated HEDI scale was used to determine the teacher effectiveness rating. The conversion scale was used to determine a HEDI value for the percentage of students who met the growth target in the SLO.

An issue with SLOs is that the individual teacher sets the goals and the goals are different across classrooms and district lines, which would lead to lack of classroom comparability (Tyler,
2011). While there is little research on SLOs, administration approval of the goal is key. Without supervision, there would be incentive for teachers to set the bar low so that students could meet or exceed the teacher’s goal (Tyler, 2011). There have been few studies on SLOs, although a Pay-For-Performance pilot program in 17 Denver schools for the Community Training and Assistance Center (as cited in Tyler, 2011) provided some insights. The study found that 89% to 93% of teachers met their SLO objectives and the quality of the SLO increased over the four-year study. A study in Austin showed that high school mathematics and science teachers who met at least one of their SLOs demonstrated greater net achievement growth than did students of teachers who did not meet their SLOs (Schmitt & Ibanez, 2011). The non-experimental nature of both the Denver and Austin studies left some doubt regarding the relationship of SLOs and student achievement and indicated additional studies on the topic are needed (Tyler, 2011).

Another problem with SLO-based evaluations is that goals have been based on individual teachers and consistency across classrooms has neither been emphasized nor required within schools in the same district or across districts in the state. At this point, it has not been clear how SLO-based evaluations can differentiate among effective teachers and more research on this topic is needed (Tyler, 2011).

Alternative Practices in Teacher Evaluation

**Instructional rounds.** Instructional rounds are an educational approach comparative to medical rounds (City, Elmore, Fiarman, & Teitel, 2009). The practice is modeled after the way physicians learn to improve their practice by visiting patients in groups and discussing treatment plans. Instructional rounds combine three elements of improvement: 1) classroom observations, 2) improvement strategies, and 3) a network of educators (City et al., 2009).

Instructional rounds rely on a detailed protocol. A small group of teachers is formed with a teacher leader. The group observes another teacher while taking notes on specific instructional strategies; this is followed by a debriefing meeting to reflect on the experience (Marzano, 2011). Instructional rounds differ from evaluation, as they are peer-to-peer and meant to improve the teaching community as a collective entity thus improving the individual teacher (City et al., 2009). Instructional rounds focus on teachers learning from each other by observing and creating discussions to better teacher practice. Widener (2014) found that teacher self-efficacy was positively impacted after participating in instructional rounds.

The goal of professional development is to produce more effective teachers (Tournaki, Lyublinskaya, & Carolan, 2011). Instructional rounds have allowed for teachers to fulfill their need to be self-reflecting, life-long learners (City et al., 2009). A study conducted with departmental leaders found instructional rounds to be supportive of teacher collaboration while developing skills that fostered professional development (Chew, 2013). These skills included direct conversational benefits with colleagues, increased expansion of their own personal skillsets, and enhancement of observational techniques and self-reflection (Chew, 2013).

**Teacher self-reflection.** Teacher self-reflection can be described as voluntarily engaging in purposeful evaluation about instructional beliefs and pedagogical knowledge, as well as the impacts of those beliefs. Shulman (1986) described reflection as a way to generate new knowledge about teaching, merging content knowledge gained from experience with pre-service pedagogy. Studies have shown that teacher reflection is the best method for building professional competence (Darling-Hammond & Richardson, 2009). Teacher self-reflection has been closely
associated with professional development, efficacy, and improved teaching practices (Sparks-Langer, Simmons, Pasch, Colton, & Sarkko, 1990; York-Barr, Sommers, Ghere, & Montie, 2001). A study conducted on pre-service teachers through the Collaboration for the Improvement of Teacher Education, an organization that promotes reflective teaching practices, concluded that teachers who engaged in reflection applied pedagogical principles when making teaching decisions (Sparks-Langer et al., 1990). When reflective teacher practice was implemented into professional development, teachers developed self-knowledge during their professional careers (Klein, 2008; Leitch & Day, 2000). Gutierrez (2015) indicated that the context for professional development must be tailored for teachers to enact skills and knowledge gained from the training.

Summary

NCLB shifted the focus of teacher evaluations from subjective observation towards a student-achievement driven model (Matula, 2011), while Race to the Top encouraged state legislators to pursue policy changes to current teacher evaluation systems (Sawchuk, 2011). One interpretation of the more recent situation regarding the use of student achievement data in evaluations is that the system is not designed to promote sound teaching practices but to secure Race to the Top funds (Copella, 2013). Some researchers have questioned the federal government establishing a public education system that endorses uncertain practices of evaluating teachers by incentivizing states to participate in order to receive grant funding.

There are clear challenges facing new teacher evaluations, particularly those in secondary science because of content specialization. Teacher and administrator perceptions of these challenges provide insights into policy changes. Potential solutions will require continued research on VAMs and SLOs, as well as including additional measures as part of teacher evaluations. SLOs need to be studied to determine how effective they are in differentiating between highly effective and effective teachers.

An effective, comprehensive evaluation system would give teachers direction and feedback, supportive leadership, and positive working conditions. It would allow schools to remove consistently ineffective teachers through valid quantification of teacher quality. Successful implementation of teacher evaluations needs to include teacher and administrator insights in order for the system to gain teacher acceptance and foster pedagogical growth.
Chapter 3

Methods

The review of literature regarding teacher evaluations is considerable, however, the literature is limited with regards to teacher perceptions of evaluations and is minimal for the perspectives of science teachers and administrators. The literature provided insights into the motivational theories that were used in the design of APPR and the associated challenges with incorporating intrinsic and extrinsic incentives. The policy implementation factors of *clarity* and *practicality* described by Fullan (2001) and the motivational theories behind teacher evaluations described by Firestone (2014) provided the framework for this research study. The methodology used in this study is outlined in this chapter.

Introduction

The purpose of this study was to understand teacher and administrator perceptions of teacher evaluation policy in suburban New York State to help guide future evaluation reform efforts. Teacher and administrator perceptions were gathered through a qualitative method of inquiry using surveys and one-on-one interviews.

This chapter discusses the methods used to collect data and analyze the research questions. The chapter is divided into twelve parts, starting with the research questions. The second part of the chapter describes the context of the study, Suffolk County, New York – a unique area to study because it possesses the characteristics of a larger school district in a suburban geographical area. The third section provides detailed information about the New York State science Regents courses and exams, specifically in two secondary science content areas – living environment and chemistry. Annual Professional Performance Reviews (APPRs) in New York are described in the fourth section. Within this section, the methods and logistics for evaluating science teachers are documented. The middle sections outline the research design, which used a phenomenological paradigm to explore the shared meaning of the participants about their experiences with APPR. Within these sections, the pilot study, protocol formation, and subjects are discussed. Finally, within the last three sections, the development of the coding scheme, methods used to analyze the qualitative data, and methods used to reduce researcher bias are outlined.

Research Questions

The purpose of this study was to examine the state of the current Annual Professional Performance Reviews (APPRs) for Regents science teachers in Suffolk County, New York. The study also investigated teacher and administrator perceptions of the evaluation system and identified some unintended consequences of APPR in Suffolk County, New York. The research questions are as follows:
1. How do secondary science teachers view the APPR evaluation system in terms of clarity, practicality, and motivation?

2. How do secondary science administrators’ view the APPR evaluation system in terms of clarity, practicality, and motivation?

3. What are the impacts of the APPR system on teachers and administrators?
   a. What challenges have been identified?
   b. What are the recommendations of science teachers and administrators on how to evaluate science teachers?

The study included current secondary living environment teachers with varied ranges of experience and administrators responsible for their evaluations across Suffolk County, New York. Participants were chosen based on their years of experience, school district demographics, and position as a secondary living environment or chemistry teacher or school administrator/supervisor. The goal of the participant selection process was to gather teachers with a range of educational experience among different school districts in order obtain a representative view of teacher and administrator perceptions. A recruitment letter was sent to teachers throughout Suffolk County utilizing maximum variation sampling techniques (Patton, 1990). The use of such purposeful sampling allowed the researcher to describe key themes that were consistent among a variety of school contexts. Consequently, qualitative data revealed unique descriptions of teacher and administrator perspectives, as well as shared patterns elicited from a heterogeneous sample (Patton, 1990). This study was approved by the Stony Brook University Institutional Review Board (#635740-2).

**Study Context**

Suffolk County is a suburban county located in the State of New York. Suffolk County is comprised of the eastern end of Long Island, measuring 2,373 square miles with 69 different school districts. Of the 19.4 million people living in New York, 1.5 million people live in Suffolk County (U.S. Census Bureau, 2010). Suffolk County encapsulates the characteristic features of a much larger school district in a small area and has a long tradition of state-mandated test-based accountability.

Long Island school districts are socioeconomically diverse. The historical development of Long Island has resulted in local school districts that vary in size, race, and income (Long Island Index, 2009). They are largely funded by local taxes on property and businesses, making each district different. Property taxes on Long Island are 60% higher than Fairfax Virginia, an area of comparable wealth (Long Island Index, 2009). The difference is Fairfax is one district that serves the entire county. There is a great range of variance in school income, expenditures and educational results because of the segregation, the fractured structure of education, and the major differences in property taxes (Long Island Index, 2009). Thirty-six percent of Long Island districts cover less than five square miles and 75% enroll fewer than 5,500 students. Table 3 provides three examples that illustrate the wide variation among school districts in Suffolk County.
Table 3
*Select Suffolk County School District Demographics, 2011-2012 (NYSED, 2011a)*

<table>
<thead>
<tr>
<th></th>
<th>Bay Shore School District</th>
<th>Central Islip School District</th>
<th>East Islip School District</th>
</tr>
</thead>
<tbody>
<tr>
<td>Free or Reduced Lunch</td>
<td>51%</td>
<td>83%</td>
<td>11%</td>
</tr>
<tr>
<td>African American or Black</td>
<td>21%</td>
<td>25%</td>
<td>1%</td>
</tr>
<tr>
<td>Hispanic or Latino</td>
<td>36%</td>
<td>64%</td>
<td>11%</td>
</tr>
<tr>
<td>White</td>
<td>38%</td>
<td>8%</td>
<td>84%</td>
</tr>
<tr>
<td>Annual Attendance</td>
<td>96%</td>
<td>93%</td>
<td>96%</td>
</tr>
<tr>
<td>Student Suspension</td>
<td>6%</td>
<td>8%</td>
<td>1%</td>
</tr>
<tr>
<td>Avg. science class size</td>
<td>23</td>
<td>22</td>
<td>22</td>
</tr>
</tbody>
</table>

**New York State Science Regents**

In order to understand a critical component of teacher evaluation in New York State, a discussion of the state standardized testing system is in order. High school students in New York State must achieve passing scores in several Regents exams to meet the graduation requirement. A Regents exam is a standardized core subject achievement exam developed and administered by the NYSED under the supervision of the Board of Regents of the University of the State of New York. Teacher committees and disciplinary subject and testing specialists prepare the examinations, which are aligned with New York State’s Learning Standards (NYSED, 2015c).

To earn a Regents diploma, students must have passing scores of 65 or better on the following subject area tests: comprehensive English, mathematics, global history, and science (NYSED, 2015a). NYSED offers the following science Regents Exams: living environment, physical setting/chemistry, physical setting/Earth science and physical setting/physics. Students must complete three units of credit in science to graduate from high school in New York. The three units must be composed of commencement level science courses aligned with New York State Learning Standards in Mathematics, Science and Technology (NYSED, 2015a), including one course from the physical setting and living environment. The third unit may come from another physical setting or a science elective. Students must pass one Regents examination in science to earn a Regents Diploma, or two for an Advanced Regents Diploma. This study only reports interviews with teachers of the living environment and chemistry Regents courses.

**The living environment**. The living environment is an introductory biology course that follows the New York State Living Environment Core Curriculum Standards. The Regents exam for living environment is designed to measure whether students have met the course standards in areas of concentration including scientific inquiry, graphing and analysis, laboratory skills, the organization of life, biochemistry, cells, heredity and genetics, reproduction and development, homeostasis, evolution, ecology, and the impact of human activities on the environment (University of the State of New York, State Education Department, 2000a).

The living environment Regents exam is composed of five sections. Part A consists of 30 multiple-choice questions. Part B-1 consists of 13 multiple-choice questions. Part B-2 consists of approximately 12 questions, some multiple-choice questions and some free response questions. Part C consists of 20 free-response questions and Part D consists of 12 questions based on required state laboratory activities.
**Physical setting/chemistry.** The physical setting/chemistry course covers a wide variety of topics relating to the nature of matter. Topics include matter and energy, bonding, periodicity, atomic structure, kinetics, equilibrium, acids and bases, organic chemistry, and oxidation-reduction reactions. Students are provided with a reference table, which consists of formulas, a periodic table and key information (University of the State of New York, State Education Department, 2000b).

As a requirement for admission to the chemistry Regents exam, students must complete 1200 minutes of laboratory experience. The completed reports must be kept on file by the school district. The chemistry Regents exam is comprised of four sections. Part A and part B-1 contain 50 multiple-choice style questions. Part B-2 and C consist of about 35 free-response style questions.

**Annual Professional Performance Reviews in New York State**

On May 28, 2011, New York’s Senate and Assembly voted to structure teacher evaluations so 40% of the composite score was based upon student achievement, and provided $20.4 million for a new longitudinal data system (Meyer, 2011). The goal of teacher evaluations was to improve teaching (Namajhi, 2010; Taylor & Tyler, 2012). A primary feature of the law was that every school district in the state prepares and implements an APPR that began in the 2012-2013 school year. The new laws have been controversial because they implied that Race to the Top initiatives would lead to better education, making this sort of teacher evaluation more permanent. Four rating categories replaced the three-category system. Teachers and principals were classified as “highly effective, effective, developing and ineffective” also known as HEDI scales (NYSED, 2015b). Teachers were annually reviewed for performance based on a composite score of student growth, student achievement and other measures such as teacher observations. These categories are described in more detail below.

**Composite score.** In New York, effective with the 2012-2013 school year, every classroom teacher in the state received an annual Composite Effectiveness Score (CES), which is a number between 0 and 100. The CES is based on three components of evaluation: student growth, student achievement, and other measures. Based on the teachers CES, the teacher’s APPR was classified into one of four categories specified in regulations (NYSED, 2013a) of the New York State Education Commissioner (Table 4). This scale is referred to as the “HEDI scale” based on an acronym of the individual categories.

<table>
<thead>
<tr>
<th>Composite Score</th>
<th>APPR Classification Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>91-100</td>
<td>Highly Effective</td>
</tr>
<tr>
<td>75-90</td>
<td>Effective</td>
</tr>
<tr>
<td>65-74</td>
<td>Developing</td>
</tr>
<tr>
<td>0-64</td>
<td>Ineffective</td>
</tr>
</tbody>
</table>

The composite score was calculated by taking 20% student growth on state assessments, 20% locally selected measures of student achievement, and the remaining 60% was based on other
measures. The state defined four APPR classification categories according to where a teacher fell point-wise and these point values were the same for every district, as defined in Table 5. The districts, however, had the freedom of assigning different point values to HEDI scales for student growth and student achievement.

Table 5
*Teacher Evaluation Rating Categories*

<table>
<thead>
<tr>
<th>Rating Category</th>
<th>Student Growth</th>
<th>Student Achievement</th>
<th>Other Measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Highly Effective</td>
<td>Results well above district goals</td>
<td>Results are well above district adopted expectations for student learning standards</td>
<td>Overall Performance and results exceed standards</td>
</tr>
<tr>
<td>Effective</td>
<td>Results meet district set goals</td>
<td>Results meet district set goals for student learning standards</td>
<td>Overall performance meets standards</td>
</tr>
<tr>
<td>Developing</td>
<td>Results are below district set goals</td>
<td>Results are below district set goals for student learning standards</td>
<td>Overall performance and results need improvement in order to meet standards</td>
</tr>
<tr>
<td>Ineffective</td>
<td>Results are well below district set goals</td>
<td>Results are well below district set goals for student learning standards</td>
<td>Overall performance and results do not meet standards</td>
</tr>
</tbody>
</table>

**Student growth.** Student growth is a measure of the change in a student’s scores between two or more points in time (NYSED, 2015b). In order to measure student growth, objectives need to be defined to show evidence that a student has learned more science. A Student Learning Objective (SLO) is an academic goal set by an educator at the beginning of each school year for the students. The SLO must contain information on the student population, the learning content, the instructional timeframe, the assessments used to measure the goal, the baseline level of the students in the class, the expected target by the end of the course, district-based HEDI ratings, and rationale as to why a teacher chose such targets. SLOs for Regents-level courses need to use the Regents exam results as the evidence of student learning during the instructional timeframe. The living environment and chemistry courses measured student growth using a baseline exam and the state Regents exam as a summative assessment.

The teacher must first establish a baseline of student knowledge at the beginning of the year. This baseline could be a prior Regents exam, a district or regional developed “pre-test.” A district generated HEDI scale was used to determine the teacher effectiveness rating. The conversion scale was used to determine a HEDI value for the percentage of students who met the growth target in the SLO. These values were different across school districts. The HEDI percent values for two different districts with similar populations of students are shown in Figures 2 and 3. The scale ranging from 0 to 20 was developed by the state. The lower numbers are the percentages of students meeting the teacher set target. These percentages were determined by the school district. There was ambiguity with these percentages because each district created its own distribution of scores. One school district may have been more rigorous in teacher evaluation than the other, as shown in Figures 2 and 3, which represent two actual school districts on Long Island. For example, in Figure 2, a teacher with a score of 54 is rated “effective” in one school
district, while in Figure 3, the same teacher in another district would be rated “developing.” In both figures, the top numbers ranging from 0-20 were state generated while the bottom percentages of students meeting target scores were set up by each individual school district.

<table>
<thead>
<tr>
<th>HIGHLY EFFECTIVE</th>
<th>EFFECTIVE</th>
<th>DEVELOPING</th>
<th>INEFFECTIVE</th>
</tr>
</thead>
<tbody>
<tr>
<td>20 19 18</td>
<td>17 16 15</td>
<td>14 13 12 11 10 9</td>
<td>8 7 6 5 4 3 2 1 0</td>
</tr>
</tbody>
</table>

Figure 2. HEDI district “A” conversion chart.

<table>
<thead>
<tr>
<th>HIGHLY EFFECTIVE</th>
<th>EFFECTIVE</th>
<th>DEVELOPING</th>
<th>INEFFECTIVE</th>
</tr>
</thead>
<tbody>
<tr>
<td>20 19 18</td>
<td>17 16 15</td>
<td>14 13 12 11 10 9</td>
<td>8 7 6 5 4 3 2 1 0</td>
</tr>
</tbody>
</table>

Figure 3. HEDI district “B” conversion chart.

Locally selected measures of student achievement. New York State Regents exam scores were required for the Locally Selected Measure in the four major science disciplines. This measure differs from an SLO in that the teacher did not set the target for the students, rather the school district set the target in a district generated HEDI scale. The HEDI scale was typically the same as the SLO without a baseline assessment or teacher target score. For example, a school district may have used a science Regents passing score as the target. The percentage of students meeting that target was then converted to a number between 1 and 20. If 72% of a teacher’s students passed the science Regents exam in school district “A,” the teacher would receive 14 points for the locally selected measure while a teacher in school district “B” would earn 13 points for this section of the composite score. This could mean the difference between a “highly effective” and “effective” teacher. A science teacher’s rating on the locally selected measure could be a number between 1-20.

Teacher observations. Sixty percent of a teacher’s composite score was based upon other measures. A science teacher’s other measures were awarded a point value of 1-60. Other measures included classroom observations using a state-approved teacher practice rubric and/or survey tools. The school districts had some latitude when assigning the breakdown of the 60 points provided by other measures. A rubric served as a guide to assess teacher performance in the classroom. The majority of the composite score came from two or more direct teacher observations, one of which was an announced or formal observation and one of which was an unannounced or informal observation by a principal or trained administrator not from parent or student surveys. A formal observation was scheduled and usually required meetings before and after the observation with the trained administrator. A trained administrator was designated by the school’s principal or district superintendent as having senior instructional and managerial
responsibility. An informal observation was one that was not scheduled but was typically followed by a post observation discussion. The state noted that at least 31 points of the 60 possible other measure points had to come from unannounced observations (NYSED, 2015b).

The NYSED selected four acceptable teacher observation rubrics, giving school districts the opportunity to choose the rubric that best suited their needs. The participating districts used Charlotte Danielson’s Framework for Teaching (Danielson, 2008) or the New York State United Teacher’s Teacher Practice Rubric (NYSUT, 2011). The rubrics were generalized so they could be applied to all subject areas.

The Danielson framework included four domains: 1) planning and preparation, 2) the classroom environment, 3) instruction, and 4) professional responsibilities (Danielson, 2008). A series of 22 components, subdivided into 76 elements were described within each domain to articulate the knowledge, skills, and dispositions required to demonstrate teacher competence in the classroom. The 76 elements of teacher quality were further disaggregated into four performance levels: 1) unsatisfactory, 2) basic, 3) proficient, and 4) distinguished. The framework was designed to acknowledge the complexity of teaching, promote professional conversation, and provide structure for self-assessment and reflection (Danielson, 2008). The participating school districts used the revised edition of Danielson’s Framework for Teaching (2011). An example of one of these components and corresponding levels of performance is shown in Table 6. Marzano et al. (2011) commented that the framework was grounded in research and was flexible enough to be used across disciplines.

Table 6
Example from Danielson’s Framework (Danielson, 2011)

<table>
<thead>
<tr>
<th>Domain 3: Instruction</th>
<th>Unsatisfactory</th>
<th>Basic</th>
<th>Proficient</th>
<th>Distinguished</th>
</tr>
</thead>
<tbody>
<tr>
<td>Component 3c: Engaging Students in Learning</td>
<td>The learning tasks and activities are aligned with the instructional outcomes but require only minimal thinking by students, allowing most students to be passive or merely compliant. The pacing of the lesson may not provide students the time needed to be intellectually engaged.</td>
<td>The learning tasks and prompts are partially aligned with the instructional outcomes but require only minimal thinking by students, allowing most students to be passive or merely compliant. The pacing of the lesson is too slow or rushed. Few students are intellectually engaged or interested.</td>
<td>Virtually all students are intellectually engaged in challenging content, through well-designed learning tasks, and teacher scaffolding by the teacher, and fully aligned with the instructional outcomes. In addition, there is evidence of student initiation of inquiry, and student contributions to the exploration of important content. The pacing of the lesson provides students the time needed to intellectually engage with and reflect upon their learning, and to consolidate their understanding. Students may have some choice in how they complete tasks and may serve as resources for one another.</td>
<td></td>
</tr>
</tbody>
</table>
The alternative rubric used by the participating school districts was the NYSUT Teacher Practice Rubric (NYSUT, 2011). The rubric was created by the System of Teacher and Development (TED), a collaborative effort by a group of educators across New York (NYSUT, 2011). TED was created in 2009 with grant money provided by the American Federation of Teachers. The goal was to support teachers by developing improved teacher evaluations. The teacher practice rubric was divided into seven standards that were aligned with the seven New York State Teaching Standards that were adopted in 2011 (NYSED, 2011b). The seven standards were: 1) knowledge of students and student learning, 2) knowledge of content and instructional planning, 3) instructional practice, 4) learning environment, 5) assessment for student learning, 6) professional responsibilities and collaboration, and 7) professional growth (NYSED, 2011b). Each standard was broken down into different elements and each element was subdivided into indicators. An example of one of these standards, elements, indicators and corresponding levels of performance is shown in Table 7.

### Table 7
**Components of NYSUT Teacher Practice Rubric (NYSUT, 2011)**

<table>
<thead>
<tr>
<th>Ineffective</th>
<th>Developing</th>
<th>Effective</th>
<th>Highly Effective</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teacher directions and procedures are confusing to students. Teacher does not adjust explanation to meet student needs.</td>
<td>Teacher directions and procedures are clarified after initial student confusion. Teacher attempts to adjust explanations to meet student needs.</td>
<td>Teacher directions and procedures are clear to students. Teacher adjusts explanations to meet student needs.</td>
<td>Teacher directions and procedures are clear, complete, and anticipate possible student misunderstanding. Teacher adjusts explanations to meet the needs of individual students.</td>
</tr>
</tbody>
</table>

**Research Design**

The researcher employed a social constructivist lens to conduct a phenomenological research study. Social constructivism, or interpretivism (Denzin & Lincoln, 2011), is an interpretive framework used where the researcher strives to “understand the world in which they live and work” (Creswell, 2013, p. 24). The goal of the social constructivist view is to rely comprehensively on the participants to build independent meanings of their experiences (Creswell, 2013). The interview questions used in the social constructivist framework were broad to allow for the participants to construct meaning in discussions while the researcher interpreted how the participants viewed their experiences (Creswell, 2013). The social constructivist framework is apparent in phenomenological studies where individuals describe their experiences (Moustakas, 1994).

The focus of this phenomenological research study was to explore and describe the shared meaning for a group of individuals about their lived experiences. A phenomenological study concentrates on describing what the participants have in common as they experience an occurrence (Creswell, 2013). This study explored teacher and administrator perceptions of APPR,
motivations to better their teaching practice, and recommendations regarding teacher evaluation policy. In phenomenological studies, data are commonly collected through interviewing individuals who have experienced the same phenomena. Since this study relied on teachers’ and administrators’ views of the teacher evaluation system in New York, semi-structured interview protocols allowed for the questions to be broad and general so the participants could construct the meaning of the interview, allowing for more discussion (Creswell, 2013). This process utilized elements of grounded theory, where unexpected key statements from interviews were identified and coded to build a novel explanatory framework.

Qualitative research is informed by certain philosophical assumptions or views deep-rooted in the researcher (Creswell, 2013). The researcher had to acknowledge her beliefs and decide whether those beliefs should be incorporated in the research study. This study was informed by the researcher’s history as a science educator and her view of APPR. As such, she consciously reduced her bias by focusing on the subjective views of the research participants. In an attempt to approach science teacher and administrator perceptions of APPR with a deeper level of understanding, the researcher acknowledged her position as a science teacher in New York and found it necessary to bracket her experiences from the participants, whereby she consciously minimized preconceptions while collecting and analyzing data (Tufford & Newman, 2012).

The goal of this study was to examine teacher and administrator perceptions on the clarity and practicality of teacher evaluation reforms, as well as how motivational theory might incentivize teacher buy-in. In this qualitative research study, data were collected from interviews at five high schools in Suffolk County, New York, following the recent implementation of the new teacher evaluation system. The researcher conducted the interviews at the participants’ schools and in the location that was most comfortable for the teacher or administrator, typically a classroom or office. By conducting the interviews at these locations, the researcher minimized the distance between her and the participants (Guba & Lincoln, 1988), allowing for knowledge to be elicited through the subjective experiences of the participants.

Recruitment and Data Collection Procedures

Each prospective participant in this study received a recruitment letter (see Appendix C). Subjects were given a consent form before participating in the study so they were aware their participation was voluntary and their identities would remain confidential (see Appendix D). Guided by Fullan’s (2001) implementation framework of clarity and practicality, administrator and teacher volunteers were interviewed 1-2 times for approximately 45 minutes. Semi-structured interview protocols were designed based on issues related to recent reforms in New York State and the research base in teacher evaluation (see Appendices A and B). Strauss and Corbin (1990) recommended at least ten total interviews with detailed coding to build a grounded theory analysis.

This study included secondary science teachers with varied ranges of experience and administrators responsible for science teacher evaluations. Participating teachers and administrators were purposefully chosen based upon their years of experience, science content area, district demographics, and positions as secondary science teachers or administrators. The study required participants to have experienced the change in teacher evaluation law, and science teachers with more than five years of experience would have been teaching during this transitional period. Science teachers were also chosen based on the content area in which they taught. Science teachers that taught the living environment course were chosen because the class
and high stakes Regents exam was required for graduation. To broaden the perspectives of the participants and to make the study more generalizable, science teachers who taught a Regents course ending with a Regents exam not required for graduation (i.e., chemistry) were also chosen.

The interviews were semi-structured and were either audiotaped or recorded with field notes taken. The interview protocol focused on the conversation, the style was casual in effort to build rapport and establish trust with the participants. The audiotaped interviews were transcribed and analyzed using qualitative research methodologies, described in more detail in the next section. Data were coded and stored in a secure location.

**Pilot and Protocol Formation**

In order to increase the robustness of the interview protocol, two pilot interviews were conducted in the winter of 2016. A school administrator and a science teacher from the same school district agreed to participate in pilot interviews. The pilot participants were asked to provide feedback and insights about the interview questions and the manner of the interviewer. Based on their feedback, the researcher adjusted the style of interview from direct questioning to more conversational. Additional probing questions were added to the interview protocols after the pilot study revealed the initial protocols contained broad and general questions and more specific topics needed to be explored. Motivational questioning was added to the interview protocol (Appendix A) to gather more information as to what drove the science teachers to improve science educational practice. Administrators were asked to comment on the impacts of APPR related to time constraints (Appendix B).

**Subjects and Context**

This study included five secondary science teachers with varied years of experience and the administrators responsible for their evaluations. When conducting phenomenological research, Polkinghorne (1989) recommended that the researcher interview 5 to 25 individuals who have experienced the phenomena. Ten participants were interviewed based on research related to subject saturation and variability (Guest, Bunce, & Johnson, 2006). After one set of interviews in the pilot study and ten additional interviews, no additional themes were found that would warrant collecting additional data. Subjects were chosen based on their years of experience, science content areas, district demographics, and positions as secondary science teachers or administrators. The participating pairs, with each pair including a teacher and the administrator responsible for her evaluation, were employed by the same school district. Since it was the administrator’s responsibility to explain the evaluation process to the teachers, these pairs were purposely chosen to analyze the relationship between the evaluator perceptions and the teacher perceptions.

To maximize subject variation, the school districts were chosen with different ethnic and socioeconomic profiles. The district demographic information was retrieved from the New York State Report Card database (NYSED, 2015d). All teachers interviewed taught Regents living environment and/or chemistry courses. The participants were all highly qualified and experienced educators who demonstrated sustained commitment to their careers. The participating teachers had professional experience ranging from 10 to 20 years. Three teachers taught living environment and two taught chemistry. Living environment teachers were chosen since all students need to take the living environment class and pass the high stakes living environment Regents exam to fulfill the New York State graduation requirement (NYSED,
Therefore, they all were evaluated in part by how their students performed on this exam. Two chemistry teachers were also interviewed to establish a range of teacher perspectives. The chemistry course and exam were typically taken less frequently to fulfill the physical science graduation requirement in New York. The New York State report card for the 2015-16 academic year indicated that 155,283 students took Regents Earth science, 106,408 students took Regents chemistry, and only 47,917 students took Regents physics (NYSED, 2015d). These numbers indicated that most students fulfilled the physical science graduation requirement by taking Earth science. Although the chemistry Regents exam did not necessarily have high stakes for the student, the student’s performance was also used to evaluate the teacher’s effectiveness. The number of science teachers within the school district and the total number of science teachers that shared the same certification as the participant were also reported. All of the participants had a least five colleagues that shared their certification. The teacher and district descriptions are summarized in Table 8.
Table 8
Teacher Descriptions and Respective School Characteristics

<table>
<thead>
<tr>
<th>Participant</th>
<th>Jack</th>
<th>Annie</th>
<th>Christine</th>
<th>Robert</th>
<th>Sarah</th>
</tr>
</thead>
<tbody>
<tr>
<td>Undergraduate Degree</td>
<td>Genetics</td>
<td>Biology</td>
<td>Environmental Studies</td>
<td>Biology</td>
<td>Chemistry</td>
</tr>
<tr>
<td>Certifications</td>
<td>Earth Science</td>
<td>Biology</td>
<td>Biology</td>
<td>Biology</td>
<td>Chemistry</td>
</tr>
<tr>
<td></td>
<td>Biology</td>
<td>General Science</td>
<td>General Science</td>
<td>Chemistry</td>
<td>Biology</td>
</tr>
<tr>
<td></td>
<td>General Science</td>
<td></td>
<td></td>
<td>General Science</td>
<td>General Science</td>
</tr>
<tr>
<td>Content Taught</td>
<td>Living Environment</td>
<td>Living Environment</td>
<td>Living Environment</td>
<td>Chemistry</td>
<td>Chemistry</td>
</tr>
<tr>
<td>Years Experience</td>
<td>18</td>
<td>10</td>
<td>16</td>
<td>20</td>
<td>10</td>
</tr>
<tr>
<td>Number of Colleagues in Content Area</td>
<td>6</td>
<td>17</td>
<td>8</td>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>Number of Science Teachers</td>
<td>23</td>
<td>41</td>
<td>30</td>
<td>27</td>
<td>38</td>
</tr>
<tr>
<td>School Demographics</td>
<td>51% Free &amp; reduced lunch</td>
<td>57% Free &amp; reduced lunch</td>
<td>50% Free &amp; reduced lunch</td>
<td>12% Free &amp; reduced lunch</td>
<td>9% Free and reduced lunch</td>
</tr>
<tr>
<td></td>
<td>53% Economically disadvantaged</td>
<td>59% Economically disadvantaged</td>
<td>53% Economically disadvantaged</td>
<td>13% Economically disadvantaged</td>
<td>8% Economically disadvantaged</td>
</tr>
<tr>
<td></td>
<td>7% ELL*</td>
<td>5% ELL*</td>
<td>11% ELL*</td>
<td>2% ELL*</td>
<td>1% ELL*</td>
</tr>
<tr>
<td></td>
<td>60% Underrepresented minorities</td>
<td>41% Underrepresented minorities</td>
<td>42% Underrepresented minorities</td>
<td>9% Underrepresented minorities</td>
<td>25% Underrepresented minorities</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*English Language Learners
The administrators’ total experience ranged from 17 years to 36 years, which included teaching experience before becoming supervisors. All of the participants had a variety of procedures for teacher evaluation and were active in the field during the evaluation policy transition. The administrators reported a range of 40 to 78 teachers they were directly responsible for supervising. Most of the administrators had undergraduate majors and certifications in science except for one. The two chemistry teachers that participated in this study were observed and evaluated by administrators who did not have chemistry majors or certifications. The administrator descriptions and district demographics are summarized in Table 9.
Table 9
Administrator Descriptions and Respective School Characteristics

<table>
<thead>
<tr>
<th>Participant</th>
<th>Jane</th>
<th>Adam</th>
<th>Charles</th>
<th>Rich</th>
<th>Stacy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Undergraduate Degree</td>
<td>Science – Med Tech</td>
<td>Elementary Education</td>
<td>Biology</td>
<td>Physics</td>
<td>Environmental Science</td>
</tr>
<tr>
<td>Certifications</td>
<td>Biology</td>
<td>Kindergarten and grades 1-6 Administrator</td>
<td>Biology General Science Administration</td>
<td>Physics General Science Math 7-12 Administrator</td>
<td>Earth Science Biology General Science Administration</td>
</tr>
<tr>
<td>Type of Administrator</td>
<td>Science &amp; Technology Director</td>
<td>Assistant Principal</td>
<td>Science &amp; Technology Director</td>
<td>Science &amp; Technology Chairperson</td>
<td>Science Director</td>
</tr>
<tr>
<td>Years Experience</td>
<td>36</td>
<td>19</td>
<td>28</td>
<td>17</td>
<td>24</td>
</tr>
<tr>
<td>Teachers Directly Supervised</td>
<td>78</td>
<td>50-60</td>
<td>40-50</td>
<td>49</td>
<td>42</td>
</tr>
<tr>
<td>School Demographics</td>
<td>51% Free &amp; reduced lunch 53% Economically disadvantaged 7% ELL* 60% Underrepresented minorities</td>
<td>57% Free &amp; reduced lunch 59% Economically disadvantaged 5% ELL* 41% Underrepresented minorities</td>
<td>50% Free &amp; reduced lunch 53% Economically disadvantaged 11% ELL* 42% Underrepresented minorities</td>
<td>12%Free &amp; reduced lunch 13% Economically disadvantaged 2% ELL* 9% Underrepresented minorities</td>
<td>9% Free and reduced lunch 8% Economically disadvantaged 1% ELL* 25% Underrepresented minorities</td>
</tr>
</tbody>
</table>

*English Language Learners
Contextual Descriptions of School Districts

School district 1. Jack and Jane were from an ethnically and financially diverse school district. The graduation rate was 89% with 94% of graduates earning a Regents level diploma. The total population of students enrolled in grades K-12 was 5,879. The district had 88% of the students earn a passing score on the living environment Regents exam, with 39% of the students achieving the mastery level score (85% or higher). A little more than half of students took the chemistry Regents \((n = 270)\) exam in comparison to those who took the living environment exam \((n = 457)\). Jack and Jane’s school district had a passing percentage of 88% for this exam yet the mastery level was 24%.

Jack seemed comfortable talking about his administrator. Jane had an extensive background in teaching science and felt her staff was comfortable with the observation dialogues and expectations. Jack described Jane as a “compassionate, tireless and effective employer” with clearly defined observational expectations. Jack and Jane shared the same biology teaching certification. The relationship was one of mutual respect and was evident in the manner in which Jack spoke of his director.

School district 2. Annie and Adam were also from an ethnically diverse school district. The total number of students enrolled in grades K-12 was 8,651. The graduation rate was 75%, which was slightly below state average with 26% of the students earning a Regents level diploma. The school district had a passing rate on the living environment exam of 68% and a mastery level of 26%. The number of students taking the chemistry Regents \((n = 210)\) dropped significantly compared to the number of students who took the living environment exam \((n = 1036)\). The passing rate for chemistry was 75% and the mastery level was 8%.

Annie and Adam’s relationship was unique in comparison to the other participants. Annie was a living environment teacher and was evaluated by an administrator with no science background. Adam asserted the importance for instructional supervision commenting, “Supervision instruction should be a huge time commitment.” Annie showed frustration with the observation process unlike the other participating science teachers, claiming she was not given specific feedback on how to raise her observation score. Adam, although passionate about developing teachers, was not a trained science teacher. He aimed to provide professional development that targeted what he believed the teachers needed, treating all teachers, science and non-science, the same, rather than focusing on specific pedagogical content knowledge for the science teacher’s discipline.

School district 3. Christine and Charles worked in a district similar to the previous two. The number of students enrolled in the K-12 district was 7,551. The graduation rate was 84% with 36% earning a Regents level diploma. The living environment Regents exam was taken by 623 students with 88% passing and 30% achieving the mastery level. The chemistry Regents exam was taken by 461 students with 72% passing and 17% mastery.

Christine and Charles demonstrated a professional relationship founded on mutual respect and trust. Charles shared with his science teachers that he believed one classroom observation was not a representation of their “entire classroom teaching experience.” Christine acknowledged that her administrator was helpful in providing productive feedback and reported positive experiences during the observation process.
School district 4. Robert and Rich were employed in an affluent school district with little ethnic diversity. The number of students enrolled in the district was 5,581. The graduation rate was 94% with 70% of the students earning a Regents level diploma. For the students who took the living environment exam \((n = 613)\), 96% earned a passing score of 65 and 59% of students mastered the exam. Of the students who took the chemistry exam \((n = 450)\), 96% passed and 44% earned a mastery score.

Each participant spoke highly of the other and their relationship demonstrated professionalism. They both characterized their dialogues as constructive, and they shared the goal of improving science instruction within their district. Robert and Rich did not share the same disciplinary teaching certification. Robert and Rich believed that teacher development was more important than a teacher evaluation score to improve science education and student learning.

School district 5. Sarah and Stacy were also from an affluent school district in Suffolk County, having the lowest percentage of students receiving free and/or reduced lunch. The number of students enrolled within in grades K-12 was 6,528. The graduation rate was 97% with 73% of the students earning a Regents level diploma. The school had 94% of the students pass the living environment Regents exam and 49% of the students earned a mastery score \((n = 597)\). Of the participating school districts, their school had the greatest percentage of students taking the chemistry Regents \((n = 516)\), with 90% of the students passing the exam and 45% of the students earning a mastery level score.

Stacy showed support for her 42 science teachers and the district “worked very hard to make sure that APPR did not get in the way of teachers wanting to continue to develop professionally.” Stacy spoke highly of her teachers, stating, “our teachers have demonstrated over and over again that they’re progressive.” Sarah pointed out inequities in terms of being evaluated by her administrator stating, “There are different levels of expectation for different teachers.”

Development of Coding Scheme

Provisional coding was utilized in this study based upon prior research synthesized in the literature review. The provisional coding process applied a predetermined set of codes that were anticipated categories or responses from the participants based on the existing literature, the research questions, and the pilot study (Miles & Huberman, 1994). This study followed Creswell’s (2013) description of provisional coding by identifying four major codes related to motivation, clarity, practicality and impression; these a priori codes were expanded into multiple categories and then combined to develop major themes in generating theory about the APPR science teacher evaluation system.

The main considerations in teacher evaluation reform are to determine how much weight should be given to evaluation data in order to reward or penalize teachers and how the evaluation itself may be used to improve instruction, thereby maximizing student achievement. These questions were informed by two different theories of motivation – extrinsic and intrinsic. The research questions were generated to determine types of incentives teachers and administrators found the most meaningful in their careers and how the APPR system leveraged motivation to promote teacher growth.

Fullan (2001) suggested that teachers change their pedagogical practices and beliefs about education when reform efforts are clearly explained and are practical to implement. Based on these constructs, teachers were asked to explain how their evaluation scores were generated.
Science teachers and administrators were also probed during the interviews to discuss their feelings about the evaluation process and how it affected their teaching practices or administrative responsibilities to assess APPR practicality.

For this study, the provisional list of codes was generated using motivational theories applied to the design of teacher evaluation policy discussed by Firestone (2014) and the two reform implementation factors generated by Fullan (2001). Based on the provisional codes gathered from the research questions, versus coding was used to identify binary terms that were in direct conflict with each other (Saldaña, 2013). Using the theoretical framework to generate the interview questions, the transcripts were coded to identify key words or phrases that fit the eight versus codes: *intrinsic or extrinsic* motivation, *clear or unclear* about the evaluation process, *practical or impractical* to implement, and a *fair or unfair* general impression.

**Motivation.** Motivation is the reason or desire a person has for acting or behaving a certain way. Guided by the first and second research questions, teachers and administrators were asked to describe what motivated them in their careers. The transcripts were scanned for key words and phrases that described the participants’ motivations as either intrinsic or extrinsic. Words or phrases like “inspiring” or “preparing students,” “personal gratification,” and “making a difference in student lives” were all common responses among the participants and could be considered intrinsic motivators. Extrinsic motivation was marked when teachers or administrators mentioned phrases such as “money,” “career ladder,” or “pay for additional activities.”

**Clarity.** Clarity addressed how explicitly policy goals were functional (Fullan, 2001). The participants were probed during the interviews to explain how their composite evaluation scores were generated. The interview transcripts were scanned for words or phrases that characterized the participants’ understanding of the process as *clear or unclear*. Since there were three major areas to gather points within the teachers’ evaluation – the two observations, the local achievement measure and the student learning objectives – the researcher looked for *clarity* both as a whole and regarding individual evaluation components. The researcher coded segments of transcript as being *unclear* about procedures when participants voiced “confusion,” “uncertainty,” or “misunderstanding” about the evaluation processes. The participants’ responses were coded as having *clarity* if they could explicitly outline the evaluation process or components of the evaluation process.

**Practicality.** Practicality refers to whether teachers considered the new policy feasible to implement (Fullan, 2001). Using the constructs of Doyle and Ponder (1977), transcripts were coded looking for words or phrases that showed *instrumentality, congruence,* and *cost* of implementing the new evaluation process. Segments of the transcripts were coded as *practical* to implement if the participants received “productive feedback” from observers that could be used to “improve practice.” Transcripts were coded as *impractical* if the participants talked about “time allocation,” lack of “productive feedback,” “slow feedback,” or not receiving “useful student data.”

**General impressions.** Finally, the interview data were coded looking for teacher and administrator general impressions of the overall system. Guided by the last research question, versus codes of *fair* and *unfair* were used to code words or phrases from the transcripts that
described how the participants felt about the evaluation process. Participants discussed areas of the evaluation system that they thought were *fair or unfair* was well as the entire process. Teachers and administrators were also probed during interviews to share their opinions of how the evaluation process could be improved and discussed their recommendations as to how science teachers should be evaluated. These descriptive phrases and the provisional coding scheme are outlined in Table 10.

Table 10
*Coding Scheme*

<table>
<thead>
<tr>
<th>Provisional Codes</th>
<th>Versus Codes</th>
<th>Descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Motivation</strong></td>
<td>Intrinsc</td>
<td>Social/emotional, reflection</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Altruistic, inspire global citizen</td>
</tr>
<tr>
<td></td>
<td>Extrinsic</td>
<td>Support from administration</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Money fear/stick</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Observed by respected peer, student growth</td>
</tr>
<tr>
<td><strong>Clarity</strong></td>
<td>Clear</td>
<td>Observation rubrics and process</td>
</tr>
<tr>
<td></td>
<td>Unclear</td>
<td>Pretest random,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Unclear metrics</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ill-conceived/implementation</td>
</tr>
<tr>
<td><strong>Practicality</strong></td>
<td>Practical</td>
<td>If observations are used for PD</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Helpful leadership, reflection to improve practice</td>
</tr>
<tr>
<td></td>
<td>Impractical</td>
<td>Admin: number of observations</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Maybe not in content area</td>
</tr>
<tr>
<td><strong>Impression</strong></td>
<td>Fair</td>
<td>Autonomy, Union Representation</td>
</tr>
<tr>
<td></td>
<td>Unfair</td>
<td>Student characteristic vary,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pre-test random, attribution to weak leadership</td>
</tr>
</tbody>
</table>

**Qualitative Data Analysis**

*Coding.* This study followed a modified *grounded theory* approach to collecting and categorizing interview data to formulate an explanatory framework (Glaser & Strauss, 1967). Rather than collecting data and quantifying it through a traditional deductive approach, grounded theory compares and classifies data based on its own characteristics. The classification criterion among the interview data was determined as the study unfolded. Different fragments of qualitative interview data were assigned *codes*, or phrases that captured essential meanings (Saldaña, 2013). Once data were coded they were placed into various categories or groups and it was possible to draw theoretically relevant conclusions (Glaser & Strauss, 1967).

Grounded theory has clear guidelines from which to build descriptive frameworks that show relationships among concepts (Charmaz, 2011). Four different stages of coding were used to categorize data: open, axial, selective and theoretical. Analysis began by examining the interview
data to establish a general sense of the information collected (Glaser & Strauss, 1967). Open coding organized the questions and responses into initial categories and common themes (Strauss & Corbin, 1990). Interview transcripts were coded line-by-line following the guidelines of Corbin and Strauss (2014), Glaser (1978), and Charmaz (2014). These initial, open codes were comparative and tentative, yet conceptualized and situated within existing research in teacher evaluation. Charmaz’s description of line-by-line coding was particularly instructive:

> Line-by-line coding, the initial grounded theory coding with gerunds, is a heuristic device to bring the researcher into the data, interact with them, and study each fragment of them. This type of coding helps to define implicit meanings and actions, gives researchers directions to explore, spurs making comparisons between data, and suggests emergent links between processes to pursue and check (Charmaz, 2014, p. 121).

The open coding process took subjects’ statements at face value and assign descriptors for further analysis. It was important for the researcher to maintain qualitative rigor by actively minimizing personal bias while interpreting subjective statements, a process known as bracketing (Fischer, 2009). Subsequent coding phases searched for causal explanations and theoretical constructs in the data (Miles & Huberman, 1994).

Axial coding reorganized the data to reveal new categories. Concepts generated in open coding were reevaluated and moved to different or more concentrated categories. When data are organized through axial coding, descriptive codes are systematically placed into specific categories, and emerging links among these categories are identified (Charmaz, 2014). Throughout the coding process, member checking provided necessary feedback for proper coding categories. Member checking is a quality control process used by the researcher to improve accuracy and validity on the recorded information from an interview; it involves presenting qualitative analyses to the subject to confirm interpretation (Coffey & Atkinson, 1996).

After axial coding, in which data were placed in preliminary categories, the process of selective coding was employed. Selective coding created a story from emerging categories by linking information between and across discrete categories. The features of various paradigms were identified and configured into connecting strands (Pandit, 1996). The integration of categories preceded the final development of a new theoretical framework.

Finally, theoretical coding aligned the data with what is known in the field. The purpose of grounded theory is to approach data inductively, that is, to generate hypotheses from qualitative information (Glaser & Strauss, 1967). Theoretical codes specify relationships, integrate participants’ perspectives on common experiences, and allow construction of a linear narrative to provide thematic insights. The propositions generated in this iterative process provisionally generalized relationships (Pandit, 1996). By providing connections between the data collected by the researcher and the researcher’s knowledge of the field, the data built upon the literature in a meaningful way (McCann & Clark, 2004).

Validity and Reliability of Qualitative Data

Standardized criteria to ensure rigor in qualitative research have been used to make the results of this study more credible (Barusch, Gringeri & George, 2011). To increase the credibility of the participants’ responses, the researcher was clear about the nature of the study
and engaged in light conversation before the interview to gain trust with the participants (Patton, 1990). Creswell (2013) presented eight significant strategies for establishing objectivity and suggested that qualitative studies use at least two of these strategies. The eight strategies are 1) prolonged engagement and persistent observation, 2) triangulation, 3) peer review, 4) negative case analysis, 5) reflexivity, 6) member checking, 7) thick description, and 8) external audits. The researcher employed five of the eight strategies within this qualitative study: reflexivity, prolonged engagement, analyst triangulation, member checking, and has provided a thick description of the study.

An important consideration in qualitative research is for the researcher to keep objectivity when collecting and interpreting interview data. The researcher must keep personal experiences in check. As a secondary science teacher in Suffolk County, the researcher in this study practiced reflexivity. Reflexivity is when the researcher actively engages in critical self-reflection about her potential bias; this is necessary since the researcher was the primary “instrument” for data analysis (Watt, 2007).

The quality of a researcher’s study is dependent upon the author’s accountability in articulating the research methods (Barusch, et al., 2011). Prolonged engagement and persistent observation helped the researcher to identify preconceptions and bracket those preconceptions from the data analysis. Lincoln and Guba (1985) explained that the researcher should work to identify the components of a situation that are most significant to the issue being followed and focus on those components in detail. The study began well before the interview process had begun. In 2015, the researcher was actively engaged studying the different APPR plans across the area and worked to understand APPR law changes. The researcher spent a year interviewing the participants and remained engaged after the interviews were complete. It was important for the researcher to identify her preconceptions of APPR and focus on the participant’s responses to questions about clarity, practicality, and motivation.

 Analyst triangulation was also employed to minimize the impact of researcher bias. This type of triangulation involves two or more researchers analyzing the same qualitative data set, or interview transcripts, and the analysts then compare their findings (Patton, 1999). To insure interrater reliability, an additional coder was given the a priori category descriptions to match with corresponding interview text. Interpretations were made individually before the researchers conferred. Comparisons between coding schemes were made and disagreements resolved through extended discussions.

 Member checking was also utilized as means to check the researchers interpretations of what the participant implied during the analysis of the interview transcripts. Member checking is a research technique used to reduce researcher bias, when the researcher confirms her interpretation of what the participant said during an interview with the participant (Barusch et al., 2011). Doyle (2007) discussed how member checking is used as a means to validate the trustworthiness of the qualitative results. The researcher asked clarifying questions during the interview discussions and contacted the participants during the analysis of the post-interview transcripts to verify perceptions and reduce ambiguities.

 The researcher provided a detailed description of the participants, their school districts, and the relationships between the science teachers and administrators. Lincoln and Guba (1985) described thick description as a method to improve external validity for a qualitative study. Holloway (1997) characterized thick description as a detailed depiction of field experiences in which the researcher makes clear the context of a situation. It is the researcher’s job to both describe and interpret observable behaviors within a particular context of a qualitative study.
(Ponterotto, 2006). The study described the research area in depth and the researcher used long quotes to give a “voice” to the participants.

**Summary**

This qualitative study examined the links among science teacher and administrator insights, the reform implementation factors of clarity and practicality, motivational incentives, and the consequences of teacher evaluation systems. This research study used semi-structured interviews to determine teacher/administrator perceptions of the evaluation reforms in Suffolk County, New York. The research identified criteria to analyze teacher and administrator perceptions to inform educational policy. Using a phenomenological approach and elements of grounded theory as a mechanism for coding qualitative data, as well as consistently comparing and organizing the data into categories, theories were generated from non-quantifiable concepts.

To insure a rigorous qualitative study, the researcher used strategies suggested by Creswell (2013). These strategies included reflexivity, prolonged engagement, analyst triangulation, and member checking. Within this chapter the researcher also provided a thick description of the participants and the school demographics in which they worked. The contributions of this research study will guide future studies in teacher evaluation reform and teacher/administrator perceptions of educational reform.
Chapter 4

Results and Findings

The previous chapter presented the context of this study and provided more detailed information about the components of APPR. The methodology chosen for this study followed a phenomenological approach, as the researcher formed a general understanding of science teachers’ and administrators’ perceptions as they experienced the APPR system. The theoretical framework guided the interview questions and the transcripts were coded to identify key words or phrases that fit the provisional codes for clarity, practicality, motivation and the participants’ general impression of APPR. This chapter discusses these results and is organized by research question.

Introduction

The purpose of this study is to analyze data from science teachers and administrators to direct future teacher evaluation reform efforts in New York State. Science teacher and administrator perceptions were gathered through a qualitative method of inquiry using one-on-one interviews. This study also explored the potential unintended consequences of the current New York State APPR system for science teachers and administrators. Four themes were evident from the analysis of the interview data and pertained to observations, time management, professional development, and student standardized test scores.

This chapter is divided into three major sections. The first segment presents the main findings of this study and is organized by research question. Each research question ends with a summary of findings. Secondly, there is a discussion of within district perspectives, or how the teacher-administrator pairs of participants perceived the APPR system. Finally, summaries of noteworthy findings are discussed in the last section. The research questions were as follows:

1. How do secondary science teachers view the APPR evaluation system in terms of clarity, practicality, and motivation?
2. How do secondary science administrators’ view the APPR evaluation system in terms of clarity, practicality, and motivation?
3. What are the impacts of the APPR system on teachers and administrators?
   a. What challenges have been identified?
   b. What are the recommendations of science teachers and administrators on how to evaluate science teachers?
Results

There were clear challenges facing the implementation of the APPR teacher evaluation system. The secondary science teachers in this study were particularly impacted because of content specialization and the high stakes exams their students were required to take. Teacher and administrator perceptions of these challenges provided insights into reform implementation and its potential effectiveness. Several major themes were evident from the interviews. The findings in this section are organized by research question.

Research Question 1
How do secondary science teachers view the APPR evaluation system in terms of clarity, practicality, and motivation?

The evaluation of the first research question revealed three major themes relating to secondary science teacher perceptions of their evaluation in New York State: clarity, practicality, and motivation. Each subsection discusses the interview results pertaining to each theme in research question one. The major themes from analysis of the interview data revealed that teachers perceived their APPR evaluations in similar ways.

From each theme, key points of interest were identified. In terms of clarity, the observation process was the most explicitly outlined component of the teacher’s evaluation. The participating science teachers were often unclear how student test scores were being calculated towards their composite score and how the local achievement measure and student learning objectives differed. In terms of practicality, the science teachers found the dialogue generated during the observation process to be the most practical. The participating teachers showed no dissatisfaction with the observation rubric, however, the baseline exams used to generate the growth score were often described as arbitrary and unreliable. Finally, in terms of motivation, the science teachers all described intrinsic means of motivation as the driving force in their teaching careers.

The first research question was answered by looking through the lens of the science teacher. The results pertaining to teacher clarity are discussed first.

Teacher clarity. Clarity addressed how explicitly policy goals were explained for these objectives to put into practice (Fullan, 2001). As Fullan (2001) discussed, teachers need to have clarity about new policies in order to feel comfortable changing their teaching practices and beliefs. This subsection identifies teachers’ perceptions of the clarity of the overall evaluation score or composite score, as well as the clarity of each component contributing to the teachers’ composite. The three components used to calculate the science teachers’ composite scores were the observations, the student learning objectives (SLOs), and the local achievement measure (LAMs). Science teachers were asked to explain the process by which they were evaluated.

The research participants generally demonstrated reasonable knowledge of the observation process and identified the rubric used to generate their observation scores. For example, Christine provided a detailed description of the process and shared that it had been beneficial for her teaching practice:

Christine: …we have a pre-conference, we have the observation, and then we have the post-conference. I come to the pre-conference with my plan, and we discuss it, and they suggest changes I make. They ask me why do you do this or do you think this might help or they'll suggest things, and I'll take those into
consideration when I'm actually finalizing my lesson plan for the observation. Then afterwards it’s like that next step how do you think it went, what could you have done differently. That's everything – there is a dialogue. I do feel that so far it's been helpful.

Christine revealed a detailed understanding of the observation procedures, explaining the pre-observation meeting as a discussion with the observer, the classroom observation itself, followed by a post conversation. The expectations were made explicit by her administration. She found the dialogue useful and productive, which was an important aspect of her buying into the process. The participating teachers were all able to identify the rubric for evaluating their lessons, the number of observations per year, and the weighting percentage of the observations in their overall composite score.

Although the teachers knew the observation process, they expressed some degree of uncertainty regarding the difference between the SLO (student learning objective) and the LAM (local achievement measure). When asked to explain how they were evaluated, the teachers often demonstrated confusion as to how student test scores were being used as a part of their overall evaluation score. Christine could not differentiate between the classroom SLO, which was based upon her own designated target to measure student growth, and the LAM, which was a district wide testing goal set by the local administration:

Christine: I have the observations, and then that growth score from my students based on the end of the year, their Regent scores – how they've grown. Then there's another component. I don't think it's either the state or the local, and get very confused on this so I'm really sorry. This is the part that I don't always – I'm not always sure of. But it’s component-based on I think something with the school. I don't even – I don't know. So, I hope that's not a problem, but I really am not fully sure about that.

This teacher revealed confusion as to how student test scores were used as a part of her evaluation, so it appears she did not understand how her actual rating was generated. She could not describe the student learning component nor the local achievement measure. Most of the teachers revealed similar confusion. Christine was unclear on how student test scores were used as a part of her evaluation and this may have hindered her from changing her teaching practices and “buying in” to the new policy.

Some of the participating science teachers were aware their overall evaluation was comprised of three pieces but could not produce the name of each component. Jack was able to clearly describe the components that constituted his composite score:

Jack: So I guess component one would be the achievement – one of the components, I don’t know their order, would be the achievement of my students on their exams. Their improvement level on the exams, how well they improved from an unfounded test earlier in the year. We use a Regents as our indicator. Second would be my observation and how my supervisor felt I performed on my observation. And third would be I guess my culture in the building, what I am involved in, what extracurricular activities I do, what things I do outside of the classroom to help benefit the overall learning environment of the building.
Jack was asked in during a member check about what he meant when he used the term “unfounded test.” He described this term as a pre-test or baseline exam. When Jack was questioned about the difference between the achievement on exams and the improvement on exams, he seemed unclear and wondered what the expectations were for the testing components.

**Jack:** That’s one of the areas I’ve actually always felt as though it was very muddled. I don’t see much of a difference there at all and I don’t understand what the expectations of the state or the expectations of my district are.

The fact that Jack could describe the three components of his evaluation but was unsure about the expectations and the difference between the SLO and LAM attests that he did not fully understand how his composite score was generated. For teachers to buy into policy changes, they must be clear on how the policy works to change their teaching beliefs and adjust their practices accordingly.

All of the teachers demonstrated confusion when it came to the way the SLO and LAM scores were calculated. They did not know how they earned points for these two components to establish their overall yearly evaluation score. Robert, an accomplished teacher, was able to identify the components of his evaluation but was unclear how the SLO and LAM scores were calculated as a part of his overall composite score. He described entering his grades into an online tracking system.

**Robert:** We enter our own grades into our e-school system and then it’s reported as a baseline in the system. Beyond that, once we enter in our Regents scores at the end of the year, the computer does its magical calculations.

Robert could describe the pieces of his evaluation but did not know how he earned his score for these two components. He, like the other teachers in this study, demonstrated lack of clarity when discussing two major components of their evaluations.

The interview data revealed that the science teachers were lacking in clarity regarding the formulation of their composite scores. They showed the most clarity when discussing the observation process, and the most confusion regarding the SLO and the LAM, reflecting a limited understanding of the APPR. The next segment addresses the perceptions of science teachers regarding the practicality or usefulness of APPR.

**Teacher practicality.** *Practicality* refers to whether teachers considered the new policy feasible to implement (Fullan, 2001). This subsection addresses the teachers’ perceptions of practicality for the overall evaluation process and each individual component of the composite score. The data indicated the science teachers recognized the importance of evaluation and felt the observation process was the most practical of the three APPR categories. The participating science teachers generally felt the dialogues between them and their administrators were practical and constructive. The baseline exams used to generate growth scores were found impractical when the exam was arbitrarily chosen and unreliable.

When asked about the observation process and its place in the rating process, some science teachers felt it had improved their instruction while others felt the administration was overworked and gave no productive feedback. Christine felt as if the conversation during a post
conference with her administrator, Charles, was productive and provided her with information that could improve her practice.

Christine: They give constructive criticism, it's never back shaming – in my experience. I can only talk about myself. But I've had only positive feedback. I've had positive experiences where things were suggested, but never anything negative, and never for sure about anything that – where I was being put down for anything I did or – they've been helpful. I haven't had any negatives with that, at least from the people that have observed me.

Christine continued to discuss how she received helpful suggestions from Charles during her preconference meeting that were potentially useful for the upcoming lesson. She described the procedure for the observation process as productive and she enjoyed the collaborative planning with Charles and other administrators who had observed her in the past. Similar to most teachers in the study, Christine felt that the observation component of her evaluation was practical because it provided her with meaningful information and suggestions to develop her teaching practice.

Jack had similar feelings about the practicality pertaining to observation rubric. He felt that he could use the rubric as a guide to help him design more effective science lessons.

Jack: Well, I feel as though the rubric does help me address certain areas that we may have overlooked. Because we do have a rubric in front of us I do have – as far as my observation – I do understand what my supervisor or the person observing me is looking for and I have it written down. Some of it to me is a little bit repetitive, some of the areas where we’re graded and each individual section or component of the rubric seems to be a little repetitive but at least I do have a document that I can keep in front of me and make sure I can touch or hit each one of those components.

The expectation of the observation was clear when a grading rubric was employed and available to teachers. Jack commented that the rubric gave him explicit, realistic guidelines for a successful lesson. Instrumentality is a crucial dimension within the practicality factor of implementing changes in educational policy. The construct of *instrumentality* refers to how the policy provides teachers with the means and the process for change (Doyle & Ponder, 1977). The observation rubric served as the means to facilitate change in the observation process. Jack felt comfortable with the expectations of the observation lesson, unlike the lack of clarity he described with the SLO and LAM. He found the rubric and the observation process practical because he was given the resources to succeed. The rubric was the instrument for translating the goals of the evaluation process into practice.

Science teachers generally expressed that the observation process was important in evaluation and development, however, they believed that using student test scores as a means of evaluating teacher performance should be eliminated or adjusted. Jack commented that he was in favor of accountability and professional improvement, however, the use of student performance data was problematic:
Jack: Certainly it’s important to be observed. The observation process is important, I believe. In any job you should be observed on a yearly level and always critiqued and you should be learning from those critiques. So certainly that is an area I think you need to eliminate or somehow adjust for using student performances to evaluate teachers’ performances.

Jack pointed out some equity issues with the APPR, since he taught the living environment course to a group of English language learners (ELLs). His students were held to the same standards as matriculated English speaking students in terms of Regents exam performance, and his evaluation score was partly based upon their achievement. He found this part of evaluation troublesome because the performance of non-traditional students was not considered fairly when the evaluation system was changed.

Jack: The curriculum is far too wide for ESL students to begin with and then that’s never taken into account. And then the fact that I felt as though I was going to be scored negatively on it I felt as though I had to rush through a curriculum and I’m sure that led to my students really not enjoying the class and not even comprehending things as well as they should have. From that perspective that was really, really difficult for me and after the first year of APPR I step back, I re-evaluated myself and I went back to some old strategies.

Jack claimed that this system was not benefiting ELLs nor was it fair to their teachers. In his opinion, the curriculum was too difficult for most of them to achieve a high score on the Regents exam. In turn, he received a lower composite score based solely upon the language barriers of his students. Jack noted that he changed his practice for the wrong reasons and quickly reverted to older strategies that he felt his students found more useful. Jack was not receiving practical information from his evaluation score to improve his teacher practice with ELLs, which limited his respect and appreciation for the APPR process.

Some teachers voiced concerns about the practicality of the observation process when they felt they were not getting productive feedback from observers. Annie commented that the delay time between the observation and written feedback was too long for the evaluation to be useful. She also felt there was no clarity regarding what could have been improved in the lesson. In this sense, she could not understand how the observer differentiated between “effective” and “highly effective” and felt “short changed” to receive a lower rating when her performance was discussed in the post observation meeting:

Annie: …my observation, I didn't get back for weeks later, even though I know it's supposed to be 48 hours… and I feel like they're just – the administration is so overworked with doing observations that I think they write just the basic stuff that they think it's almost like a form letter. But yet, it's just like the old way of observations… so I got an "effective," but yet there was nothing in there that you would've improved about the lesson.

This teacher was not only disappointed with the excessive time to receive feedback, but also the cursory manner in which the administrator used the rubric as a list of items or behaviors to verify. She believed that the performance measures should have been more explicit and the observer
should have exerted greater effort in evaluating her work to produce insightful instructional guidance. Annie felt that she could not raise her evaluation rating from “effective” to “highly effective" since the administrator did not provide a pathway for her to do so.

Based on the conversations with the science teachers, the practicality of the SLO component was also called into question. Annie expressed problems with the SLOs created for her students by the school district. Her baseline score for the SLO component was not a science exam but the average student score on the New York State English Language Assessment (ELA):

*Annie:* …so many students opted out of the ELA this year so that our – what was suggested, which is so statistically wrong, is that we take the average of our students who did take it in that class and the average becomes the SLO or the baseline for every other student who opted out, which doesn't even make sense. Because we have no idea if the students who actually took it are on the high end of the spectrum, middle, low ELL students. We have no idea. So to take an average of students just because they took it doesn't really apply to the students who didn't take it.

Of the five teachers interviewed, Annie’s school district was the only district that did not use a science baseline exam. Rather, the baseline score for her students was an average of all students that took the ELA and was correlated to students’ predicted scores on the living environment Regents exam. Annie did not know how the scores were correlated. This called into question the validity of the measure, how clearly this component was explained, and the practicality of the SLO component in her evaluation. The SLO was supposed to measure the teacher’s influence on student growth. This evaluation component becomes impractical if it does not measure the teacher’s impact on the students’ science knowledge. Annie did not gain useful information to improve her science teaching practice since her SLO score was based on the ELA:

*Annie:* The baseline for our students is based on the seventh grade ELA exam. And then that somehow is miraculously converted into a score that correlates to a Regents exam score. And that's the – that is the score that the students have to achieve for me to – for me to be in that "effective" range or into those ranges.

Christine’s school district did something similar to compute baseline scores. Her district did not create a baseline living environment pretest but instead used the students’ scores on the eighth grade state standardized Intermediate-Level Science Test (ILS). She also pointed out the impracticality of the SLO:

*Christine:* …because it's living environment, they typically look at the eighth grade tests, the ILS given in eighth grade in the middle school, and they have some sort of system... create a growth score or what the student is expected to achieve on the Regents in my class. Each student let's say has their own score. So, I could have kids that are only expected to get a 60 even though 65 is passing, because their growth is such that they did so poorly in eighth grade that they – this is what they’re expecting if they go over even better, which a lot of kids can do and will do.
Christine explained that students who scored low on the ILS may only have to score 60% on the living environment Regents exam and that score apparently showed growth within her class. The target for growth was set low and most of the students met or exceeded the target. Christine’s response called into question the validity and practicality of the SLO. She was not evaluated on the growth of her students in the subject she was teaching but with an exam the students had taken two years prior. Christine did not set the target for improvement for her students within her content area and did not gather any information about the prior knowledge her students possessed. In Christine’s case, she was not eliciting practical information from her SLO data to improve her teaching practice and the arbitrary nature of the process cast doubt upon its fairness and effectiveness.

The other participating science teachers reported using a content specific baseline pre-test as a part of the SLO. The baseline exam is a pre-test given at the beginning of the school year. These exams are usually in the content of the course and are designed by the teacher or administrator. The questions were influenced by older versions of the specific content Regents exams. Robert reported that his students randomly chose answers on the pre-test, which he claimed was not an appropriate way to measure the student’s actual growth. Robert described how some students created patterns of answers on the bubble answer sheets, which called into question the validity of the baseline exam:

*Robert:* Yeah but it’s so flawed because that SLO exam, I mean, if they’re guessing to making patterns as you know, and so their scores are coming back at what, five, ten, out of 100 points maybe 25 at most, and so it’s not the appropriate way to measure growth of a student over time.

The pre-test based upon previous chemistry Regents exams did not reveal much since the students had little knowledge of the chemistry content assessed. Students were aware that baseline exam was used to evaluate the teacher and did not affect their grades; consequently, the students were less serious about the pre-test. Robert believed the pretest scores were completely random. Sarah explained similar experiences with her students, claiming that the students did not take the baseline seriously because they take this type of pretest in every Regents level course:

*Sarah:* Oh, they don’t even take it seriously. Yeah, they have to take it in all their Regent’s classes, so they’re burnt out. And they know that they don’t know it. And also our administrator and our superintendent came out and he’s like against all the testing and everything. So he’s like, we’re not pressuring the kids to try as hard as they can. They usually get anywhere from five to 30% correct.

The students’ growth in content knowledge could not be determined reliably when the students filled in pre-test answer sheets randomly or were not encouraged to do their best. Teachers need to have confidence in the validity of the process to buy into policy changes to adjust their teaching beliefs and practices.

Science teachers found that APPR was flawed in some ways and was a somewhat unfair method of evaluating teacher quality. A major flaw that was exposed from the interview data dealt with the unfair nature of student test scores as a part of a science teacher’s evaluation. The last segment of research question one dealt with science teacher motivation. Science teachers were asked to describe the motivation supporting their teaching practices.
**Teacher motivation.** The design of teacher evaluation policy in New York has been influenced by extrinsic and intrinsic motivation theories to reward great teachers, remediate ineffective teachers, and develop sound teaching practices. With these motivational incentives in mind, science teachers were asked to comment on what motivated them in their careers.

All of the teachers interviewed were motivated intrinsically. They reported self-reflection, the desire to be a better educator, passion for content, and satisfaction in making a difference in the lives of their students. Robert shared that his professional motivation came from two main sources – his sincere interest in his discipline and learning new content, and the rewarding nature of working with high school students:

*Robert:* I think it’s the passion for the subject in terms of what I’m teaching. I’ve taken such an interest in chemistry. I’ve also taken such an interest in learning new things through the science research program. The second part is the interaction with the students and just working with kids – students – that interaction is incredible. It’s fun and it’s the reason I like coming here every day.

The conversation about Robert’s motivation was insightful. Robert’s resume in education was impressive and his passion for the subject was echoed in his teaching practice. APPR did not provide him with supportive policies that kept his intrinsic motivation strong. The policy was not congruent with his practices and beliefs. He suggested that the APPR did not nurture his motivation, rather the extrinsic incentives attached to APPR clashed with his science teaching beliefs.

Some teachers in this study voiced concerns with extrinsic motivation, stating that it would lead to teacher competition rather than teacher collaboration. One of the goals of APPR was to develop excellent teachers, however, categorical rating comparisons to within-school and out-of-school colleagues had consequences. When external evaluation promotes competitiveness and ranks teachers in relation to their colleagues, these same teachers may compete for a higher composite score and demonstrate unwillingness to share effective pedagogical tools. Jack described this potential threat to the spirit of professional collaboration:

*Jack:* Colleagues should really work together. Part of teaching, part of the idea of education is mentoring, learning from other individuals, sharing knowledge and I feel as though when you put that competition aspect into it you’re not going to inspire people to be better, you’re going to inspire people to kind of be greedy and money always leads to that and when you bring that variable into an equation, who knows where it goes, who knows how it’s going to fracture education.

Jack believed that when money is used to inspire teachers, they might potentially hoard resources, which would reduce teacher collaboration. Teachers may be less willing to share best practices if extrinsic incentives are attached to teacher performance. Jack believed continuous collaboration is an essential part of education. This view is important because when designing evaluation policy, extrinsic and intrinsic incentives have been difficult to assimilate (Firestone, 2014), and educators have expressed suspicions about the consequences of such evaluations. For Jack, monetary incentives could damage relationships and hinder the very essence of teaching as a process that is learned through best practices of other educators.
Christine discussed the unfairness of tiered compensation models due to the wide variation in student quality year-to-year:

*Christine:* I don't think that's a good thing to have because it's not fair. Just because I'm highly effective and somebody else is not, I'm getting more money, but what if I got better kids that year. Because after 16 years I've had some great years, and then there's other years where I've had lower level students or students that are not performing as well as other… those scores or whatever they're using to get people more money are not indicative of the person's actual ability, and level of teaching.

Christine pointed out the inequity between classes and between different school years. Some teachers taught higher-level students when students were tracked. Students were placed in different levels, based on academic ability, within the same content course. An example of this would be an enriched living environment course, taught with additional topics not tested on the exam. The teacher who has higher-level students might typically receive a higher composite score. She believed that money should not be used as an incentive because teachers do not pick their students; some years, students are better than others. Because of this inequity, she stated that the composite score should not be used to judge teaching performance.

Sarah and Annie felt differently about extrinsic incentives than the other science teachers, however, they did agree as to what personally motivated them in their career was intrinsically based. Annie believed that “there does need to be some accountability for your students’ progress.” Sarah commented on the importance of holding teachers accountable. She felt that the state exam was a way to hold the teachers accountable and to motivate them to care about student outcomes. She compared a chemistry class that ended with a Regents exam to an introduction to chemistry class that did not:

*Sarah:* If the teachers aren’t held accountable for something, there’s no like okay, you have to do this, then they’re not going to teach the kids. Like the reality is, the teachers are, there’s still this you want the kids to be able to do well on the Regents so you have to teach all this. In our other course, we have another course where it’s intro chem. It’s where there’s no state exam. They don’t feel the motivation. They’re like ah, whatever, you gotta do what you gotta do. And there’s not as much pressure and they really don’t do as much. And they don’t care about the kids as much. They’re not as motivated to try new things to get better. Unfortunately, because of the unions and this protection, you have to do something to make some of the teachers accountable. There’s too many teachers that just don’t care.

Sarah blamed collective bargaining and union protection for creating a system that did not hold all teachers accountable. She believed that the state test at the end of the term was a form of extrinsic motivation that pushed Regents level teachers to care how students performed and held them accountable for what was taught. Sarah brought up an interesting issue with the collective bargaining units across each district. Each district had its own teacher union, which helped negotiate the terms for APPR implementation. She claimed that the teachers’ union protected
teachers that “just don’t care” about experimenting with new educational practices or teaching students the science content.

**Summary.** Overall, the participating science teachers commented that intrinsic motivation was a driving force in their success. Even with student test scores as a key component of their evaluations, which all agreed was not the best way to evaluate teacher quality, the internal rewards of teaching were more influential and outweighed the external pressures of a high composite score. The participating science teachers were in agreement that adding extrinsic motivation, such as pay for performance, could damage the very essence of teacher collaboration.

Key points of interest were generated from the themes of clarity, practicality, and motivation. In terms of clarity, the science teachers were able to clearly describe the observation process but showed lack of clarity when describing the difference between the SLO and LAM. The science teachers found the observation process to be the most practical and useful in improving teaching practice. The baseline exams used to generate growth scores were impractical and unfair when they were arbitrary, invalid, and unreliable.

To answer the second research question, the lens was changed from science teacher to science administrator. The results for research question two are discussed in the next section.

**Research Question 2**

*How do secondary science administrators’ view the APPR evaluation system in terms of clarity, practicality, and motivation?*

The second research question was informed by the administrators’ perceptions of science teacher yearly evaluation in terms of clarity and practicality. The administrators in this study were responsible for implementing APPR by evaluating the participating science teachers. Their perceptions were important as they influenced the teacher’s perceptions regarding the clarity and practicality of the APPR system.

Within the two themes of clarity and practicality, specific points of interest were discovered. In terms of clarity, the administrators found the timing of the implementation of the new policy to be poor. They reported that the rushed timeline diminished the teachers’ clarity. Each administrator showed a clear understanding of the components that comprised the science teacher’s composite evaluation score. In terms of practicality, three key points of interest were identified. The amount of time required to perform and document the observations was impractical for administrators. Consequently, many did not buy into the reform effort as the state intended. They also reported that conversations between administrators and science teachers had improved as a result of the observation process. Finally, the administrators agreed that two observations per year was not a true representation of teacher effectiveness.

The results discussed below look through the eyes of science administrators and how they perceive APPR. The results for the interview data pertaining to administrator clarity are addressed in the first segment.

**Administrator clarity.** The data revealed that the timing for the implementation of new evaluation policies in New York led to some confusion. As APPR was introduced, Common Core Standards (Common Core State Standards Initiative, 2010) were implemented simultaneously. The Common Core Standards were new to teachers and some Regents curricula
changed. The Common Core did not affect high school science teachers directly but contributed to the uncertainty of teacher evaluations.

Administrators were asked to comment about the rollout of APPR. Adam described how the transition in teacher evaluations was poorly timed and did not allow for teachers to fully understand the process:

Adam: The rollout was a problem statewide because it came out when Common Core was first introduced. And there was a lot of research that clearly outlined the fact that you had so many initiatives under the Race to the Top – because APPR was part of the Race to the Top – when you secured those funds, then you had to implement or jump through all these hoops. And at the same time, Common Core, we had the shift in the standards. And, because in New York State, we were actually tested on the new standards and then evaluated a year before we had to do it. So there was definitely a poor timeline as far as implementation. There wasn't enough time for the teachers to understand the new standards in order to be evaluated correctly.

Adam suggested that politics drove the speedy implementation of teacher evaluation law changes in New York as the state was trying to secure funds offered by Race to the Top. He also suggested that teachers and administrators should have been included in the discussion about meaningful teacher evaluations. Adam explained that his science teachers had lack of clarity about the new evaluation policy because the timeframe to implement the change was too aggressive. Clarity was crucial for teachers to embrace educational policy changes.

New York State received approximately $700 million as a part of the Race to the Top educational grant in 2010. Administrators participating in the study brought to light unintended consequential issues with the evaluation process. Adam suggested that money and politics were the driving mechanisms for the changes in the teacher evaluation law. The quick timeline to implement the change left no time for discussion of how teachers should be evaluated:

Adam: Race to the Top was a federal initiative and it allowed school districts or states to clamor for any money that they could get. And – so in my opinion, it was really the politics that really got in the way, in addition to probably the time and not having enough teachers and administrators have input into that, to how teachers really should be evaluated.

Adam believed that state efforts to receive funds offered by Race to the Top minimized valuable discussions on how teachers should be evaluated. He suggested that science teachers resisted the APPR policy because they were not included in the preliminary discussions. However, all of the administrators showed clarity discussing each component of the science teachers’ composite scores. The next analysis of research question two addresses the views of the science administrator on the practicality of APPR.

Administrator practicality. Science administrators were asked to comment on the time spent performing the duties of APPR and they voiced concerns over the commitment required for more than 40 formal observations per year, which included a pre-observation meeting, the observation, and a post observation conference. The time involved in the observation process
was significant in terms of practicality, and they preferred devoting attention towards staff
development and improvement rather than the typically rote and cursory observation process.
Charles described the increase in workload on science administrators and how it shifted his
priorities:

Charles: Everybody is very sympathetic towards the teachers, and rightfully so;
however, it's the administrators that are carrying the weight… you have to do all
those observations. Managing time became very critical, and it kind of took away
from things that you'd normally do.

Charles stated that the observation process was lengthy and consumed the majority of his time,
which he felt could be used more efficiently in providing mechanisms for the science teachers to
improve practice. Cost was a key consideration in the implementation factor for practicality. The
amount of time required to perform the observation and pre- and post-conference meetings was
impractical for administrators to buy into implementing the reform effort.

The increased workload was a common theme among the administrators. As a result of
APPR, most school districts required one announced or formal observation and one unannounced
or informal observation for tenured teachers. Rich discussed how the informal observation was
never “technically” unannounced:

Rich: In addition to that there is informal observation that is technically
unannounced. I interject those informals as – to make them as convenient as I can
just because of the sheer volume. So, I try to interject them in between formals,
and therefore I will give a heads up the week before saying I may be in and
around and about.

When Rich referred to “informals,” he was stating that the observation that was supposed to be
informal was actually announced. The purpose of the informal observation was to observe the
science teacher without the planning that typically goes into a formal observation. The informal
observation was a way to see science classroom interactions day-to-day. The purpose of the
informal observation was nullified if teachers were aware an observer was going to visit. The
informal observations were not practical to implement as the state intended due to the time
constraints of administrators.

The theme of workload was not only noted in the time to perform a large number of
observations, but in the administrator’s time to document a teacher’s performance. Jane was
asked to comment on the time commitment involved in evaluating her teachers:

Jane: It's really increased with the workload. For the formal observations, you
have the pre-observations, you have the observation, the post-observation, and
then you have the write-up. The write-up is very time consuming. And then the
unannounced you also have obviously the observation. Now you're meeting after
for a full period at least and then the write-up. So I think with someone who has
a large staff like myself it's very difficult to get through. I think I do 78
observations.
Jane addressed the increase in paperwork she was required to complete for the teachers’ professional files. Jane was personally responsible for performing 78 observations per year. The “write-up” was the documentation that was provided to science teachers after the observation was complete and became a part of their professional records. She lamented the time needed to meet with the teacher for the pre-conference, observation, and post-conference, as well as time needed to professionally document the observation in a meaningful way.

The administrators often made modifications to adjust for time. Stacy discussed the option for teachers to perform, what she called a “professional growth plan” in lieu of a formal observation:

*Stacy:* A professional growth plan was developed by the teacher, where they would choose an area that they wanted to research and show growth in, and they would develop a project or a series of lessons, or something around that idea in conjunction with the administrator. So we would have a pre-[conference], they would undertake their project. There was a lot of self-monitoring involved in that. The administrator would come in at some point for a 15-minute observation of the PGP in action in their classroom. And then there was always a post-observation conference.

“Professional growth plans” were not listed on the district’s APPR policy. This modification to the APPR saved Stacy time, as her teachers did not require a full formal observation. Stacy explained that formal observations took her 45 minutes to write and professional growth plans only took 15-20 minutes. The time cost was too high for some administrators to find the evaluation process practical, influencing their ability to implement as APPR intended.

The science teachers that participated in this study felt the observation process was an important part of evaluating and developing teachers. The participating administrators agreed that the observation was essential in developing their staff, however, two observations per year did not paint an accurate picture of teacher performance. Rich pointed out the importance for instructional supervision and the time needed to support his educators effectively:

*Rich:* Supervision instruction should be a huge time commitment, but it should be equitable and fair. So many times, you're spending all this time on a system that isn't, or in my opinion, fairly evaluating the teacher. In the NYSUT rubric, the "effective" and "highly effective" categories sometimes are not attainable in a lesson. And if a teacher's only getting observed formally and informally twice and that's what's going to serve as their basis for their 50 points, then that's a problem.

The evaluation process outlined in APPR was not congruent with Rich’s professional beliefs regarding instructional supervision. He felt that teachers were receiving evaluation scores based upon categories in a rubric that were not possible in every lesson. He described the unfairness of assigning a score based on two lessons out of countless lessons over the school year. This was often not a true representation of the teacher’s ability.

Stacy had similar feelings about the formal observation process outlined in APPR. She believed that the formal observation did not provide her with means to support her science teaching staff nor did it provide information needed for her to improve her practice. Stacy found that replacing the formal observation with a “professional growth plan” was more beneficial in targeting professional development and created more productive conversations.
All of the administrators found the dialogues between themselves and their science teachers were beneficial. All agreed that the observation process allowed for productive discussions. Stacy classified all the conversations she had with her science teachers as productive, whether they chose to do a formal observation or participated in a professional growth plan:

\[\text{Stacy: I see great value in the conversations that I have with teachers around both. So, any time I’m talking to a teacher one on one, is a valuable use of my time, no matter what it is, that’s bringing us together. Because our conversations can always evolve and don’t necessarily have to center around any one particular task.}\]

Stacy revealed, as did the other administrators, that discussions with science teachers were essential for teacher growth and development. APPR did account for teacher-administrator conversations as seen in the observation process.

Administrators found the goals of teacher evaluations did not put all science teachers on the same playing field. Since teacher quality was linked to student performance on standardized science tests, it was essential to acknowledge the variance in student population that was assigned to different teachers. Several administrators regretted there was no mechanism to adjust for these differences when looking at aggregate test scores for individual science teachers. Charles described how the results of standardized tests were not a true testament of student knowledge and learning:

\[\text{Charles: It’s a superficial show of learning, and that’s what I think these standardized tests are. Some students test very well, others don’t. Some students take the same test, and then they’re allowed modifications – they’re allowed extended time, the test is read to them, the test is recorded for them, they’re given all the time they need to take it. And then other students, because they fall within certain parameters, they don’t get that luxury; they have to do it in a certain amount of time all on their own… And I don’t think that standardized testing, or a test, should ever be the sole measure of a – it’s a good indicator – but it should not be used to evaluate the instructor in any way.}\]

Charles pointed out the inequities of evaluating a teacher with student test scores because the students take the exams under different conditions. Students may have taken the same standardized science Regents exam but may have received different testing accommodations documented on Individualized Education Programs (IEPs), legal documents afforded to students who were classified as eligible for special education. Evaluating teachers using standardized test scores was not actually standardized since some students took the exam under different conditions. Charles believed student test scores should not have influenced a teacher’s evaluation score and should never have been the only measure of teacher quality.

\[\textbf{Summary.}\] The administrators had similar feelings about APPR in terms of clarity and practicality. Within the theme of clarity, the administrators believed that the fast rollout of policy in New York resulted in teachers’ lack of confidence in APPR; they were not afforded the time to understand fully the procedures of the new policy. Two key points of interest were identified within the theme of practicality. The administrators found that time to document and observe their science teaching staff was overwhelming and adjusted components within APPR to manage
time better. They also reported that the conversations they had with the science teachers had improved and were a crucial component of teacher growth and development. The final research question is answered in the next section.

Research Question 3

What are the impacts of the APPR system on teachers and administrators?

Science teacher impressions and recommendations. One purpose of this study was to gather information on how science teachers viewed the evaluation process and its impact on science education. Science teachers were questioned about their feelings regarding the APPR system and how it impacted their teaching practices and beliefs. This section describes science teachers’ recommendations on how they envisioned these evaluations could be improved.

The interview data identified several key points of interest regarding the science teacher’s impressions of the APPR system. Generally, they found the evaluation process to be inconsistent and imbalanced. The goals of the policy did not evaluate science teachers fairly due to the variance in student ability and performance. The science teachers all agreed on the importance of being evaluated, however, they felt placing a quantitative value on their effectiveness was ambiguous and sometimes unfair.

Insightful recommendations regarding teacher evaluations were elicited from coded data. The science teachers recommended that the focus of teacher evaluations should be on significant, content specific professional development rather than an effectiveness rating. They recommended the establishment of peer review systems and self-reflection documentation to develop science teaching practices and fulfill the intended goals of the APPR system. Science teachers focused their impressions on three key aspects: 1) the importance of accountability, 2) considerations on student variability, and 3) fairness of the APPR. The science teacher recommendations are also outlined.

Importance of accountability. The science teachers agreed that evaluations were important and necessary. Annie described the importance of teacher evaluation and added her thoughts on external evaluators’ experience in the science teaching profession. She commented further about the use of student test scores as a means to evaluate a science teacher:

Annie: I think that teachers should be evaluated. I am a strong believer in that, and proponent of evaluation. And I think it should always be external. I don’t think it should be internal at all, actually – but I do think that it needs to be someone who has experience in teaching and possibly even experience teaching that subject. I don’t think that it should ever be linked to student scores and I don’t think that teacher should be associated or – and tied into a school – one school score. I do think, though, that there should be maybe a slight component as to growth, but not – but like you had said, pretest, posttest, and not some random based on it correlates to some other score. I believe that there does need to be some accountability for your students' progress.

Annie recommended that an observer who did not know the teacher personally be a part of the observation process. She also felt she would gather more useful knowledge for improving her teaching practice if the observer had the proper content knowledge to make such
recommendations. Annie and Adam’s relationship was unique in comparison to the other pairs because Adam did not have any science teaching certification or background. Annie described earlier that her observation scores quantified her as “effective” with no feedback from Adam to raise her effectiveness score. Annie believed teacher evaluation should never be linked to student test scores but that the teacher should be held accountable for student outcomes in some way. She did not offer a solution this problem with teacher accountability.

**Consideration of variations in students’ backgrounds.** Christine and Jack had similar concerns pertaining to student test scores claiming that the evaluation system was flawed due to “many variables” that were unanticipated. Both science teachers worked with ELLs and felt that the evaluation process did not take this population into account. They questioned the fairness of the evaluation process with evaluating all science teachers on the same scale. Christine described this concern:

*Christine:* …the English language learners, there's no way they're going to do as well as other people. They're coming from different countries, they're coming without education for four years, and now they're in a high school level class. They might have fourth grade education, and because of their age they're now in high school. We don't choose which students we get, not that you would want to. But it's not fair to some teachers to judge everybody on the same scale when you have very different population of students from one period even to the next in your classes.

Christine’s comment gave important insights into the evaluation system. Teachers do not get to choose the type of student they teach, however, she felt an important consideration is for teachers to educate all students without the fear of losing their jobs. She also mentioned that the population of students from period to period was different and the teacher could be considered effective with one of group of students but ineffective with another.

In Suffolk County, the districts were locally organized and funded by individual school district taxes. The participating teacher and administrator pairs came from a varying range of socioeconomic status. As the discussion with Christine continued she pointed out the unfairness among different districts financially:

*Christine:* Even from district to district there are districts that have tons of resources and money and support. Then other districts do not so you can’t even compare. It's just totally different. I don't like that whole thing where it's everybody's the same, but at the same time we do have to do something.

Christine felt she was at a disadvantage when it came to her yearly evaluation because her district did not have access to as many resources as other more affluent schools in Suffolk County. The population of students was largely segregated among school districts. Some school districts had large populations of ELLs and students that were in financial need whereas others did not. Christine claimed that not all teachers in Suffolk County taught the same kinds of students and to evaluate teachers with the same metrics across New York State was unfair.
**Fairness in conducting evaluations.** While all the participating science teachers discussed the importance of evaluation, the underlying theme of unfairness was consistently identified in one manner or another. Most of this unfairness was uncovered during the discussion of student test scores. Robert, however, described how the evaluation process was flawed because the observation process was flawed:

*Robert:* So therefore, I’d say the evaluation system is flawed in a sense that if it’s based on an observation that is set up, that’s basically scheduled and people are coached and people prepare special and put on a show, then no, it’s not a true evaluation. And two, if it’s based on scores where teachers have spent weeks on end preparing students for standardized tests, I don’t know if that’s a true evaluation of a teacher and their effectiveness. Now keep in mind we do have an unscheduled observation and a pop in, but even then, we’re given a heads up.

Robert described the observation process as a “show” and not a true testament to the teacher’s daily teaching habits. He claimed that because the observation was scheduled, the teacher could prepare a lesson that earned them a good score. He also pointed out that the informal or unannounced observation was not how it was practiced within the school districts and that teachers were aware of when an unannounced observation was going to happen. Robert questioned if there would ever be a “true” evaluation for teacher effectiveness.

When asked about her general impression of the overall teacher evaluation system, Sarah pointed out inequities within APPR among her colleagues. Sarah explained the science teachers within her district were all classified as “highly effective” which painted an inaccurate picture of the true makeup of her colleagues. She felt that teachers within her department were ineffective and not improving but still received a highly effective rating:

*Sarah:* Ineffective, and they’re rated highly effective. So I could work my butt off, and someone could come in, leave, lecture every day, do no activities, not even meet the lab requirements, they’re still highly effective. So it’s just, it’s unfair. The whole system’s unfair. I don’t feel like it achieved its goal.

Sarah was asked to elaborate what she thought the goal of APPR was.

*Sarah:* …that a child is provided with an education where the teacher is qualified. And if the teacher’s not qualified, we go to the teacher and make sure we provide training or assistance and make the teacher qualified.

Sarah suggested that the goals of APPR were not realized in her school district and teachers were not given the training to become more qualified. This, she believed, was the result of having all science teachers classified as highly effective when in actuality they were not. Sarah believed that the system of teacher evaluation “was set up poorly” and wanted to see an evaluation system that would benefit students the most and provide content specific professional development.

**Recommendations.** All participating teachers offered insights and recommendations for the evaluation process. One of the main goals of APPR and teacher evaluation policy was to provide
a pathway for professional improvement. Robert explained the idea of self-reflection documentation as a part of the teachers’ evaluation:

Robert: I said early about self-reflection, so I’d like to see something that’s – this is where I’m blanking on what his title of active research; I forget what it was called. But basically, the idea was you go through a lesson or even a unit and at the end you stop and evaluate and you say okay, what worked, what didn’t work, where am I going to make changes, where am I going to make improvements? Oh, that lab wasn’t very good, I can do something better, that demo was incredible but it didn’t work for every class, how come? It has to be that reflection on your teaching practice and that has to be documented… but there has to be that reflection piece and that has to be documented and that has to be shared with colleagues, with administrators.

He felt that every teacher should self-reflect on lessons throughout the year. He believed that adding this practice would lead to better discussions with administrators and peers and would ultimately develop better teaching practices. The same teacher also commented on creating a peer review system:

Robert: It’s almost like a peer review system would work very well where you can have teachers come in and they can observe at any point in time and you can ask them to evaluate you intently. I know that takes time, I know not everybody is comfortable doing that, you have to have a rapport with that person but I think that’s a much better evaluation system than having an independent consultant come in or an administrator who hardly knows you come in a see you in an either observed or planned observation. So that should be a part of it as well.

For Robert, having his colleagues comment on his practice would be more beneficial than having an overworked administrator. This idea seemed practical for the peer observer, as well, since they could learn new techniques to implement in their classrooms. Robert believed that adding components to the evaluation process would inform the teacher and allow her to be more self-reflective to improve her practice. Also, this strategy could be used to develop great teachers rather than scoring existing ones. He recommended that test scores and non-informative observations should not be linked to a teacher’s effectiveness and non-formative observations should changed to more meaningful evaluative measures. Robert discussed the need for the science teacher evaluative process to be focused on ways to motivate teachers intrinsically. He believed making the evaluation more of a learning experience with teacher collaboration would improve instruction:

Robert: So, get rid of the test score nonsense, get rid of the planned observations and go towards something that’s more going to intrinsically motivate you to do a better job.

Robert had some visions of what he thought the evaluation process should be. He believed the focus should be more on teacher development and a collective community effort to transform education. This vision included science teachers working together to inspire each other to try
new things. An important component of such a teacher network would be to strengthen intrinsic motivation and inspire a community of educators to grow professionally to improve student learning.

Summary. Generally, the science teachers found the evaluation process to be flawed and unfair. The science teachers pointed out that student variations in ability and performance were not considered when APPR was established and the policy provisions did not evaluate science teachers accurately. They all agreed with the importance of being evaluated, however, placing a quantitative value on their effectiveness was done somewhat arbitrarily and often diminished their professionalism.

Insightful recommendations regarding teacher evaluations were discovered throughout this study. The science teachers recommended that the focus of teacher evaluations should be on meaningful, content specific professional development. They suggested the establishment of peer review systems and teacher collaboration networks. Self-reflection documentation could improve science teaching practices and be incorporated into science teacher evaluations.

The previous sections addressed science teacher perceptions of fairness and recommendations concerning their evaluations. The following sections address science administrator impressions and suggestions for science teacher evaluations.

Administrator impressions and recommendations. The purpose of this study was to gather information to make purposeful recommendations for science teacher evaluation policy changes in the future. The administrators were questioned on their impressions regarding APPR and how it had impacted science education within their school districts. Several key administrator impressions of the APPR system were elicited. Generally, the participating administrators had similar perceptions of the APPR system to those of the science teachers. Administrators reported that the politically driven policy changes rushed the implementation, creating a stressful transition and lack of teacher trust. Finally, the administrators found the one-size fits all model of APPR did not take into account economic and demographic differences within school districts.

The participating administrators offered recommendations for improving the science teacher evaluation process. Like the science teachers, the administrators agreed with the necessity of accountability to state agencies. As discussed in the previous section, the administrators generally felt that under the current APPR requirements, time had become an issue and diminished their effectiveness as educational leaders. The recommendations offered by the administrators focused on reducing the amount of time spent evaluating science teachers and increasing the amount of time for professional development. Three insightful recommendations were elicited. First, the administrators recommended the replacement of formal observations with other means of evaluation, with renewed focus on creating meaningful professional development. Secondly, they recommended an increase in informal classroom visits, portfolios, and peer evaluations. Finally, the administrators recommended educator involvement in constructing practical science teacher evaluations and local control over implementation.

Replacing formal observations with other means of evaluation. Three of the five participating administrators recommended the substitution of formal administrator observations for a peer evaluation system. Those administrators believed by reducing the number of formal observations each year, time would be available to engage in more productive educational leadership roles. Rich explained that during negotiations for the new APPR, to be implemented
in September of 2016, members of the team discussed unique ways to implement the new mandates. He discussed how a peer-to-peer observation system could allow time for the administrators to work with the teachers more efficiently:

*Rich:* When we first started the transition to the new APPR, and there were some initial committee meetings to discuss the negotiation of the new framework, many of us, myself included, said let's start thinking outside the box. Let's get a little more creative. Let's work back to something that still satisfies these mandates, but at the same time affords us the opportunity to work more effectively with our staff, and instead of maintaining this performance based show in some regards – but let's free up that time so that perhaps we can form – I would love to institute peer-to-peer observation. That's talked about all the time, rarely implemented effectively, and certainly not in our district.

Rich felt that by implementing peer-peer observations, more time could be dedicated to professionally developing his staff. The time saved could be utilized to work on new programs to advance science education in his district. Jane agreed, stating, “peer evaluation could be very helpful.” Other administrators commented on peer reviews. Stacy described how she envisioned the implementation a peer evaluation component:

*Stacy:* I would love to put something in there that includes a peer evaluation component. Specifically though with lots of training that would include instructional rounds. And I think that that part is very important. I think that teachers should be able to consistently and frequently visit each other’s classrooms. I think that doing it as a team with administrators would promote incredible dialogue that doesn’t exist right now. But there would have to be very, very clear protocols put in place, lots of professional development before you rolled it out. And I think that that would take us from where we are to the next level.

“Instructional rounds” were inspired by medical rounds, where physicians visit patients in small groups and discuss medical problems and treatment plans. Stacy discussed how instructional rounds could be beneficial in her school district with the proper instructional training. She felt that administrators could facilitate these types of observations and that the conversations generated from such collaborations were lacking in her school district.

Some of the administrators recommended less formal observations. Time management seemed to be a key finding in the practicality of the APPR system among the administrators. The recommendations made by the participants revolved around creating more time dedicated to being an educational leader. Charles recalled that tenured teachers had other options to replace formal observations earlier in his career, and should be re-established:

*Charles:* Let me just say this. In my particular district, if you were a tenured teacher you were not observed every year. You were evaluated every year but there were options that you could choose rather than a formal classroom observation. There were other things that you could do like a collegial circle,
research project, or there were different things that you could choose to do rather than being observed. Because often times it was a dog and pony show and teachers [felt] it’s worthless: “Why even bother? I’ll do something instead of that for my yearly evaluation.”

Charles discussed that before APPR was introduced, teachers had more freedom to choose the ways they were evaluated. He believed that the observation was often a poor representation of a teacher’s effectiveness because it was based upon one scripted lesson. Charles made the assumption that the teachers being observed also believed this and would choose another form of evaluation that would give them meaningful feedback.

Others gave examples of activities that could replace formal observations. Jane recommended “portfolio work” as a means to develop more creative laboratory activities. Stacy discussed the use of professional growth plans as a means of evaluation. Charles commented on more informal visits, which was how the old system worked in his district:

*Charles*: That was fine. Because I feel, as an administrator, I would observe my teachers informally often – drop in informally, or drop in between the classrooms that were prep rooms, that I could just hang out there and listen to what was going on. Because I know my presence in the classroom is going to affect student behavior and everything like that. So I would do it informally, but if I had a problem with a teacher I would deal with it way before a formal classroom observation.

Charles, like some other administrators, found more value in frequent informal visits to the classroom. The formal classroom observation seemed to be merely a technicality for him. He found that doing informal visits gave him more information about a teacher’s ability than the “show” of teaching during a formal observation.

Many of the administrators found that although the observation component of APPR generated great discussions with their teachers, it was impractical in terms of time. Recommendations were made by the administrators that would reduced the amount of time spent formally evaluating teachers and utilized that time to educationally lead their staff. Stacy, like Charles, believed that “frequent, targeted walk-throughs” were more valuable than a formal classroom observation. Stacy was questioned about the observation component of APPR:

*Stacy*: Full period? I don’t think they have a lot of value. I really don’t. I think actually, an administrator can learn a lot more about what’s happening in a classroom by doing targeted and frequent walk-throughs, lasting no longer than five to seven minutes. Not in a formal evaluative way, but informally, really gathering qualitative data about what’s happening in the classroom, to do two things. One, to support any particular teacher in their instruction and their pedagogical growth. But two, to actually drive what the administrator finds is necessary for sustained and ongoing professional development for the department as a whole.

Stacy found value in the idea of frequently visiting a science teacher’s classroom informally. Time management was a recurring theme among administrators. Stacy believed, if afforded the
time, that more classroom visits would generate more useful and meaningful information to drive professional development opportunities. She would be able to make better recommendations for workshops or professional development classes if she could pinpoint the areas they were most needed.

Focus on professional growth with teacher input and local control. As time influenced the administrator’s recommendations, so did the focus on professional development. The administrators generally agreed that the focus of science teacher evaluations should be on growth and development rather than on an effectiveness rating. As discussed in the previous section, science teachers also agreed that professional development would better science education over teacher quality ratings. Generally, the administrators would like to have seen fewer requirements in APPR regarding student growth and achievement and more requirements for professional development.

The administrators recommended educator involvement when creating new science teacher evaluations that could be used as guidelines for districts to implement productive evaluations and science teacher development. Adam offered insight into the difficulties of having a state generated teacher evaluation:

Adam: Every teacher in the state should be evaluated the same way. Because there's a difference between the observation process and the evaluation process. So because of the collective bargaining, which is the biggest, really, divider, which really creates a wedge between school districts, it doesn't allow for districts to evaluate, even under the same umbrella of the APPR, the same, very fairly and equitably.

Adam voiced concern for a uniform teacher evaluation system across the state. He felt that the strong collective bargaining units took what was supposed to be a common system and created different evaluative procedures for each school district. He felt that the teacher unions hindered the reform effort and this resulted in an unsuccessful teacher evaluation policy. He further commented that the districts should have control over how they evaluated their teaching staff:

Adam: It's either that or you send it to the local level, like we used to, and say that we – our educational leaders are highly trained on how to evaluate instruction and teacher and we're gonna – we're going to leave it at that. I mean, we went from a very qualitative district evaluation based on all attributes – not so much test scores.

Adam felt that the educational leaders within the school district should be given professional courtesy since they know what is most effective for their own teaching staff. He believed collective bargaining would impact reform efforts if they were meant to be uniform across the state. Because these units are strong in New York, Adam believed the evaluation control should be given back to individual school districts.

Some administrators discussed turning the power over teacher evaluation to local school districts with another form of accountability. The administrators understood that public education is bound to state funding and bureaucrats need to provide proof to the public that the system is working. Charles discussed what he believed to be the ideal way to evaluate a teacher:
Charles: The ideal way to evaluate a teacher would be to have a local school district establish his or her own set of parameters. And because we are beholden to the state for money, and things like that, certainly there would have to be some measure of uniform accountability that would be acceptable to the state. Because we are a state agency, although teachers are professionals, they’re not in private practice. They are part of a bureaucratic system, and there are things that have to be accountable.

Charles discussed the need for an evaluation system because public education is bound to state funding and must prove to the public that the educational system is working. Charles believed that the school districts know what is best for their own communities. He thought district high school graduates could provide meaningful data to show the system is productive. Charles recommended an evaluation system, generated by teachers and administrators, that was “fair and equitable to the students, to the parents, to the community, and to the administrators.”

Rich also recommended giving the control of teacher evaluations over to the individual district while holding the supervising staff to evaluative measures rather than the individual teacher:

Rich: I think we have to trust that our central administrators, our district chairs, our principals who certainly should be supervised and evaluated to some leadership matrix so that we know that they’re building professional capital and development amongst teachers.

Rich talked about the level of professional trust that should exist between the state and the administrative staff of each district. He discussed how the administration should have some level of accountably to the state, and the state should trust the district professionals that they are doing the right thing with regards to evaluating their staff.

**Summary.** The administrators agreed that accountability is important, however, holding individual teachers accountable for student outcomes was not working. They recommended other forms of evaluation with a focus on professional development instead of an evaluative effectiveness score. Their recommendations involved creating more time for them to be teacher leaders rather than teacher evaluators. The administrators highlighted three main recommendations. First, they recommended decreasing the number of formal observations and replacing them with frequent informal classroom visits. Second, peer-peer observations utilizing instructional rounds were suggested. Finally, they recommended educator involvement in the creation of improved science teacher evaluation and trust in professional administrators to control evaluation details at the local level.

**Within District Perspectives**

School district 1. Jack and Jane shared similar perspectives about the APPR system within their district. They were clear about the observation process and Jack felt the observation provided him with information to better his teaching practice. The pair agreed on the importance of being evaluated and found the conversations generated during pre- and post-observations were beneficial. Jack and Jane shared the same teaching certification, which may have contributed to
the way he viewed the practicality of the observation process. Jane found the observation process to be impractical due to the amount of time she spent documenting and observing a large number of teachers. Jane demonstrated clarity when she described the LAM and the SLO, while Jack demonstrated lack of clarity in these areas. Both Jack and Jane found that the LAM and SLO were impractical for different reasons. Jack taught a population of ESL students, and felt that evaluating him based on how these students performed on an exam put him at a disadvantage. Jane felt that student test scores should be analyzed over longer periods of time not just one school year.

**School district 2.** Annie and Adam’s relationship was unique in comparison to the other participants. Annie was a living environment teacher and was evaluated by an administrator with no science background. Both Annie and Adam were able to describe the observation process and agreed that formal observations were impractical. Annie felt that because her observer did not have science content knowledge, she did not receive valuable information to improve her performance and score during conference discussions with Adam. Adam found the observation process to be impractical because he believed two observations per year was not a true representation of teacher effectiveness. A unique finding was discovered when the pair was asked to explain the SLO and LAM. Annie demonstrated clarity on both components, while Adam could not clearly describe them. The rest of the administrators were able to clearly describe each component of the composite score. Both Annie and Adam found the SLO and LAM impractical. Annie found that because the baseline exam was not a content specific exam, she did not gain useful information to improve her teaching practice. Adam found the testing components to be impractical because the policy did not account for student variability.

**School district 3.** Christine and Charles perceived APPR in similar ways. Charles was an older administrator close to retirement who openly shared his beliefs about APPR with his teaching staff. He wanted his staff to know that the APPR did not reflect their individual effectiveness but rather represented a series of “hoops” the district was forced to jump through. Christine and Charles were both able to clearly outline the observation component. Christine believed that the observation was the most practical of the three evaluation components. She felt she was receiving positive, productive feedback from her supervisor and found the rubric to be a helpful tool in the design of her lesson. Christine and Charles both enjoyed the conversations they shared in observation pre- and post-discussions. This pair did have the same teaching certifications, which could have contributed to their productive interactions. Charles found the observation component impractical. He did not believe that one announced and planned formal observation provided the evidence he needed to evaluate his staff. He also felt that tenured teachers did not need to be observed every year and the forced observations were a waste of valuable time that could have been used for professional development. Christine demonstrated lack of clarity when asked to explain the SLO and LAM and found both components impractical because they provided little information for her to use to adjust her teaching practices. Charles was able to clearly explain both the SLO and LAM but also found them impractical. He argued that student test data should not be used as an evaluative measure because of the variability in student population.

**School district 4.** Robert and Rich believed that teacher development was more important than a teacher evaluation score to improve science education and student learning. They shared
similar feelings about the productive conversations they had during pre-/post-observations. They both agreed about the importance of the observation but felt that more frequent informal classroom visits would be more beneficial. The both recommended the addition of a peer review component. Robert talked more about the diminished reliability of the SLO due to students’ lack of seriousness on the pre-exam. Robert was a passionate and experienced educator who felt that APPR did not improve the teaching of science – rather, its focus was on preparing students to do well on an exam. Robert was the only person in the study to recommend teacher self-reflection. Rich felt that valuable time was not being utilized in the correct way. He wanted to be more of a teacher leader than a gatekeeper to teacher ratings. Both participants believed the focus of evaluations should be on professional development and a collaborative district effort to enhance science education.

**School district 5.** Sarah and Stacy were able to describe clearly each of the three components that contributed to composite scores and found all three components to be impractical in evaluating science teacher effectiveness. Both participants felt the observation process generated positive dialogue, however, it could be improved with more informal classroom visits and a peer review system. Stacy recommended the use of instructional rounds, and believed it could create a dialogue between administrators and science teachers that had been lacking. Sarah believed that peer review and observation by a science teacher within the same certification would be more effective than with her administrator, even though Stacy had the same certification. Stacy commented that that APPR did not work in her high performing school district, because the science teachers “have demonstrated over and over again that they’re very progressive and tend to the whole child.” Stacy believed that the “one size fits all APPR plan” for a state as diverse as New York did not work. Sarah suggested that “ineffective” teachers were being labeled “highly effective,” and felt more differentiation was needed to reward highly skilled teachers.

**Notable Findings**

All of the participating teachers and administrators showed clarity when asked to describe the observation process of APPR. The observation process has been a long-standing component of teacher evaluations dating back to the late 1960s (Marzano et al., 2011). Teachers and administrators may have demonstrated the most clarity on this component of APPR because of this long-standing observational tradition in education. Most of the participants felt that the observation process within APPR was not practical, with the exception of two living environment teachers. These teachers voiced that they received productive feedback from their administrator. This finding was notable and may be related to the fact that an administrator that shared the same science content certification observed both Jack and Christine.

It was also notable that most teachers and administrators demonstrated clarity regarding student learning objectives and all found this component impractical in practice. A similar finding was discovered with the local achievement measure. All of the participants found this component to be unrealistic as a means to evaluate teacher effectiveness. The findings pertaining to student test scores and teacher evaluations were important, as these components have been recently added as a means to evaluate teacher effectiveness. Table 11 summarizes the science teacher and administrator perceptions of clarity and practicality for each component of APPR.
Table 11
*

Science Teacher and Administrator Clarity and Practicality Perceptions

<table>
<thead>
<tr>
<th>Evaluation Component</th>
<th>Observations</th>
<th>Student learning Objectives</th>
<th>Local Achievement Measures</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Theme</td>
<td>Clear</td>
<td>Practical</td>
</tr>
<tr>
<td>Theme</td>
<td></td>
<td>Clear</td>
<td>Practical</td>
</tr>
<tr>
<td>Jack</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Jane</td>
<td>Yes</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Annie</td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Adam</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Christine</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Charles</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Robert</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Rich</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Sarah</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Stacy</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

* Administrator

Benefits, challenges, unintended consequences, and recommendations were reported by the participants. It is notable that all participants understood and welcomed teacher accountability. The participants found that the instructional conversations between teachers and administrators had increased and reported this as benefit of APPR. The participating teachers showed no dissatisfaction with the observation rubrics, however, some described the need for more direction from administrators in terms of pedagogical content knowledge. Pedagogical content knowledge refers to the teachers understanding of what makes the learning of a specific topic easy or difficult and the preconceptions the students bring to the learning of this topic (Shulman, 1986). This was a notable finding, as the selected observation rubrics did not offer specific assessment measures for science specific pedagogical content knowledge.

Teachers and administrators reported challenges and unintended consequences as a result of APPR. Notable challenges included excessive time commitment by administrators to perform and report teacher observations, arbitrary and unreliable baseline exams, and unreasonable expectations when considering student variability. The most remarkable findings dealt with the unintended consequences of this type of evaluation system. One teacher described how APPR did not foster the teaching of science but rather the preparation of students to take standardized testing, while another voiced concern that this system was not improving science instruction.

The participants were asked to make recommendations as to how they envisioned the ideal science teacher evaluations. All of the participants believed that observations should be a component of teacher evaluation and even recommended more frequent informal observations, with the addition of peer observations and collaborative teacher networks. This finding was interesting, as the participants were all searching for ways in which the evaluation could improve their instruction rather than rate their effectiveness. Table 12 summarizes the benefits, challenges and recommendations made by the participants.
Table 12

*Science Teacher and Administrator Response Summary*

<table>
<thead>
<tr>
<th>Subject</th>
<th>Benefits</th>
<th>Challenges/Unintended Consequences</th>
<th>Recommendations</th>
</tr>
</thead>
</table>
| Jack Biology     | • Rubric served as an instrument for lesson expectations and has improved his instruction | • Unclear about testing expectations  
• Student population (i.e. ELL) plays a role in teacher evaluation score  
• Inequity in student ability and performance was not considered when APPR was put into place | • Observations should remain a part of teacher evaluations  
• More teacher collaboration as a part of evaluation - PLCs  
• Focus more on school community involvement as it is “underemphasized in the APPR process”  
• Curriculum is to difficult for non-traditional students – value added should be measured in different ways |
| Administrator     | • Important to be observed – “In any job you should be observed yearly” – views accountability as part of professional responsibility |                                                                                                   |                                                                                   |
| Jane Administrator| • Conversations about instruction improved with her science teachers  
• Observation expectations are clearly outlined in the observation rubric | • Increase of paperwork documenting observations  
• Time management – 78 educators to observe per year | • Addition of portfolios, curriculum writing, updating lab activities  
• Observations are valuable as the instructional conversations improved – teachers are receptive to comments and feedback  
• Testing “has its place” and should be analyzed over longer periods of time |
| Annie Biology     | • Believes teachers should be evaluated – views accountability as part of professional responsibility | • Evaluators with no science knowledge observing science teachers is problematic  
• Delay between observation and feedback from observer – immediacy improves changes in practice  
• No path/recommendations for raising evaluation score  
• SLO baseline not content specific – arbitrary | • Observations remain with evaluators that have science content knowledge – external observers  
• Include a mechanism for teachers to see their growth and plan improvement  
• Evaluations should never be linked to student test scores with the exception of meaningful growth scores (not arbitrary baselines) |
| Adam Assistant    | • Supervision instruction is a huge time commitment, and it should be equitable and fair | • Time commitment extensive both inside and outside of school to document and observe 60 educators  
• Pushback by teachers because of poor timing with the rollout of Common Core Standards  
• Claims teachers feel they are not | • Include teachers and administrators in discussion about teacher evaluation policy  
• Reform efforts should be from bottom up  
• More time to rollout new evaluation reforms  
• All educators across the state should be evaluated the same way with the absence of union bargaining or leave the evaluation |
| Principal         |                                                                                                   |                                                                                                   |                                                                                   |
- evaluated properly – they do not trust the advice given to pursue professional development
- Politically driven educational reform and the “clamor” for *Race to the Top* money resulted in many inequities – student variability (i.e. ELL or poverty)
- Top-down reform efforts restricted teacher buy-in

<table>
<thead>
<tr>
<th>Christine Biology teacher</th>
<th>Observations and discussions with administrators provide productive feedback for professional growth</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Finds the rubric helpful in planning her observation lesson to better her instruction</td>
</tr>
<tr>
<td></td>
<td>Unclear differentiating SLO from LAM</td>
</tr>
<tr>
<td></td>
<td>SLO baseline not content specific</td>
</tr>
<tr>
<td></td>
<td>Flawed evaluation system because of inequity in student population – ELL students have the same goals as non-ELL matriculated students</td>
</tr>
<tr>
<td></td>
<td>Feels as if pre-tests “make student feel bad” so that growth can be shown – unfair for students</td>
</tr>
<tr>
<td></td>
<td>Continue productive discussions about teaching practice</td>
</tr>
<tr>
<td></td>
<td>Alternative graduation routes for the nontraditional student</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Charles Administrator</th>
<th>Observation discussions are essential for teacher growth and development</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Danielson Rubric was “easy to use” and a “decent” rubric</td>
</tr>
<tr>
<td></td>
<td>Time management</td>
</tr>
<tr>
<td></td>
<td>Goals of APPR do not evaluate teachers fairly</td>
</tr>
<tr>
<td></td>
<td>Believes student test scores are a form of “scolism” – superficial show of learning</td>
</tr>
<tr>
<td></td>
<td>APPR was a result of bureaucrats that “speak the language of data” – systems like this do not produce educational improvement</td>
</tr>
<tr>
<td></td>
<td>More time for administrators to have productive discussions with science teachers to help teacher development</td>
</tr>
<tr>
<td></td>
<td>Less formal observations and more informal classroom visits would gather more information to drive professional development ideas.</td>
</tr>
<tr>
<td></td>
<td>Student test scores should never be used as a part of a teacher’s evaluation because of inequities among students.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Robert Chemistry teacher</th>
<th>Great dialogue with his administrator about collective efforts to improve overall science instruction within his school district</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Found written documentation of observed lesson a valuable tool for reflection – it serves as a reminder of professional improvement goals</td>
</tr>
<tr>
<td></td>
<td>Unclear on how he earned SLO and LAM evaluation scores</td>
</tr>
<tr>
<td></td>
<td>Students are less-serious about pre-test; randomized responses diminish reliability</td>
</tr>
<tr>
<td></td>
<td>Observations are a “show” of teaching ability because the process is well known and prepared for – 1 lesson out of 180</td>
</tr>
<tr>
<td></td>
<td>Self reflection teaching journals</td>
</tr>
<tr>
<td></td>
<td>Peer review more practical and beneficial feedback than overworked administration</td>
</tr>
<tr>
<td></td>
<td>The focus of evaluations should be on professional development and a collaborative district effort to enhance education rather than just rate it</td>
</tr>
<tr>
<td></td>
<td>Establish a system that intrinsically motivates</td>
</tr>
<tr>
<td>Rich Administrator</td>
<td>Sarah Chemistry Teacher</td>
</tr>
<tr>
<td>--------------------</td>
<td>-------------------------</td>
</tr>
<tr>
<td><em>Observations are important to identify areas in which teachers can improve</em>&lt;br&gt;• Observations are important and dialogue between administrator and educator are productive</td>
<td><em>Chemistry students show large amounts of growth – her rating was always highly effective.</em>&lt;br&gt;• Find the professional growth plan put in place by her district in lieu of an observation worthwhile, – Receives productive feedback from her administrator.&lt;br&gt;• Assessment scores provide information for the success of her students</td>
</tr>
<tr>
<td><em>Informal observations are not unannounced</em>&lt;br&gt;• Time management – difficult complete observations with documentation – takes up valuable educator leadership time&lt;br&gt;• Two observations are not enough to rate a teacher’s ability</td>
<td></td>
</tr>
</tbody>
</table>
Stacy
Administrator
• Finds professional growth plans as a valuable tool in developing teachers
• Conversations are of great value and use of her time
• Danielson Rubric is good for science teachers because it is “student centered”

• Formal observations do not have much value
• APPR does not work in high performing districts
• “Isn’t very clear research associated with value added types of evaluative tools”

• Targeted and frequent classroom walk-throughs
• Peer evaluations with administrator facilitator would promote “incredible dialogue that doesn’t exist right now”.
• Professional growth plans could be utilized as a part of a science teacher’s evaluation and would foster teacher collaboration.
• Instructional rounds would create dialogue that is nonexistent under the current system.
Summary

The views of both teachers and administrators are important, as they are closest to the learning process (Goe et al., 2008). This phenomenological research study strived to identify science teachers’ and administrators’ perceptions as they experienced policy changes in teacher evaluation as a result of APPR.

The motivational theories applied to the design of teacher evaluation policy discussed by Firestone (2014) and the two reform implementation factors generated by Fullan (2001) provided the framework for interview questioning. The goal of this study was to examine teacher and administrator perceptions, the clarity and practicality of teacher evaluation reforms, as well as how motivational theory might incentivize teacher improvement. In this qualitative study, data were collected from interviews at five high schools in Suffolk County, New York, following the recent implementation of the new teacher APPR system. Four reoccurring themes were evident among the interview data. The four themes pertained to observations, time, professional development, and measuring student achievement.

Notable points of interest were generated from the themes of clarity, practicality, and motivation. In terms of clarity, the research subjects were able to clearly describe the observation process and showed some lack of clarity when describing the difference between the SLO and LAM. The science teachers found the observation process to be the most practical and useful in improving teaching practice while the administrators found this component impractical because of the time dedicated to the process. The participating science teachers commented that intrinsic motivation was a driving force in their success and were in agreement that adding extrinsic motivation, such as pay for performance, could damage the very essence of teacher collaboration.

All participants made suggestions and recommendations for improving science teacher evaluations. The research subjects agreed that accountability in education was important but holding the individual teacher accountable for student growth and achievement was ineffective in their school districts. The science teachers and administrators believed that the ideal way to improve student learning was to increase professional development. The implications of these findings will be further discussed in Chapter Five.
The results gathered from interviews were presented in the previous chapter. The researcher also exposed her bias as a science teacher and took steps to minimize this bias by using techniques such as reflexivity, member checking, and analyst triangulation. The researcher also described the study context as being unique and the results may not be generalizable outside the study area in rural or urban districts.

While there are limitations associated with this study, the results of this study have laid the groundwork for future studies on science teacher evaluation. This chapter will provide a discussion of the results and provide details on the four major findings gathered from the interview data. The chapter will discuss conclusions and implications across the field of education and the limitations associated with the study. The chapter will conclude with suggestions for future studies.

Introduction

Research by professional teachers with K-12 teaching experience offers the science education research community a level of specialized knowledge and the views of these individuals should be more influential than the views of those outside the field (Lederman & Lederman, 2014). Lederman and Lederman (2015) suggested that research conducted by teachers in collaboration with university researchers is superior to research conducted by teachers or researchers alone because of the diversity of perspectives that enter the study. The authors asserted that a gap exists between science educational research and practice (Lederman & Abell, 2014).

Lederman and Abell (2014) addressed the usefulness of science education research within the practice of science teaching and learning in the preface of the Handbook of Research on Science Education. The authors claimed that many published studies in science education have not been applicable to educational practice and science education research has often not impacted teacher practice directly (Lederman & Abell, 2014). Lederman and Abell (2014) suggested that science education research be rooted in educational practice for the purpose of improving teaching and learning.

This study helps fill the gap between educational research and practice by offering research that is geared to change science teacher practice through the evaluation process. This study was intended for both educational researchers and educational practitioners. It provides data that will inform science practice within the science classroom directly and has addressed questions of educational importance. It also provides practical recommendations for science teacher evaluation policy and practice. The research subjects provided valuable information on how they perceived teacher evaluations, including the importance of teacher accountability, the need for more content specific professional development, and the questionable validity of student growth and achievement measures when applied to teacher effectiveness.
As a result of *Race to the Top*, teacher evaluations and educational reforms have received significant attention across the U.S. in recent years. National policies have overhauled local teacher evaluation systems to include multiple measures of teacher performance with ties to student test scores. These evaluations were also intended to provide routes for professional improvement. New York State changed the parameters for teacher evaluation as a result of receiving *Race to the Top* grant money and developed the APPR system. The purpose of this study was to gain an understanding of how science teachers and their administrators perceived this system and their recommendations for maximizing its usefulness towards improving pedagogical practice. A goal of the study was to gather teacher and administrator perceptions of APPR to inform future policy initiatives for the evaluation of science teachers.

Research has shown that there have been concerns about whether these new evaluation systems will result in improved instruction and increase student learning despite the numerous resources devoted to teacher evaluation. Cohen and Spillane (1992) found that the efforts to improve instruction resulted in little change and educational policy research indicated that the perceptions of the teachers were important in understanding policy success and failures (Datnow & Castellano, 2000). These studies suggested that more research is needed on teacher perceptions of evaluation systems yet there are few to date.

This research study builds upon Firestone’s (2014) work with teacher evaluations and motivation theories. Although research designed to measure teacher quality has been growing (Baker et al., 2010; Bell, 2012; Bill and Melinda Gates Foundation, 2013), little research has been dedicated to evaluating programs and improving teaching. Teacher evaluations can only improve instruction and student learning if there is trust between the teachers being evaluated and the evaluation agency (Firestone, 2014), and trust in professional relationships has been found to contribute to increased collaboration leading to improved student learning in science (Smetana, Wenner, Settlage, & McCoach, 2016). This research study has not only built upon Firestone’s work with evaluations but also with a recent study conducted by Jiang et al. (2015) on teacher perceptions on evaluation reform and Marzano’s extensive work with teacher evaluations (Marzano, 2011, 2012; Marzano et al., 2011). This research study has added to the literature regarding teacher perceptions of evaluation reforms and what has influenced those perceptions, and has laid the foundation for evaluation policy recommendations.

**Research Analysis**

This phenomenological research has laid the groundwork for making policy recommendations for science teacher evaluations. This study involved the shared experiences of science teachers and administrators as they navigated changes in teacher evaluation policy. The goal was to examine teacher and administrator perceptions, the clarity and practicality of teacher evaluation reforms, as well as how motivational theory might incentivize teacher “buy-in” to future reform efforts. This study reports the analysis of interviews with five secondary science teachers and their respective administrators responsible for science teacher evaluations.

Provisional coding was used in this study based upon prior research critiqued in the literature review. Fullan (2001) implied that teachers change their pedagogical practices and beliefs about education when reform efforts are clearly explained and are practical to implement. Based on these constructs, teachers were asked to explain how their evaluation scores were formed. For this study, the provisional list of codes was generated using motivational theories applied to the design of teacher evaluation policy discussed by Firestone (2014) and the two reform
implementation constructs generated by Fullan (2001). Using the theoretical framework to generate the interview questions, the transcripts were coded to identify key words or phrases that fit the eight versus codes: intrinsic or extrinsic motivation, clear or unclear about the evaluation process, practical or impractical to implement, and a fair or unfair general impression.

**Major Findings**

It is implied from this study that evaluation policy in New York did not have teacher “buy in” because the policy lacked clarity and practicality. Science teacher and administrator perceptions of APPR have received limited attention prior to this research. Four important themes were discovered as a result of this research study. The four findings pertained to the value of teacher observations, time constraints, collaborative professional development, and accountability measures tied to student performance.

Science teachers and administrators found the conversations about science lessons had improved as a result of APPR and the required observation rubrics. The majority of the teachers were scored using the Danielson Framework for Teaching Evaluation Instrument (Danielson, 2011) and the remaining teachers were evaluated using the NYSUT Teacher Practice Rubric (NYSUT, 2011). The teachers found that the observation rubrics provided the instrumentality or the explicit means to succeed with clear lesson expectations and received more productive feedback as the administration had more to discuss using the rubric as a lesson guide. All of the participating teachers desired more content specific recommendations. However, it is notable that the science teachers observed by administrators who did not share the same teaching certification voiced this need more often. Generally, the teachers found the observation component to be impractical and suggested peer observations as a way to gain more content specific feedback.

The recurring theme of time was evident among the administrators. Administrators found the observations to be important, however, the increased volume of observations reduced the amount of time they could dedicate to being teacher leaders. The recommendations made by the administrators focused on reducing the amount of time associated with formal observations. The administrators believed the formal observation provided a limited view of the teacher’s overall practice; the science teachers also mentioned formal observations being a “show” of teaching. The administrators believed increasing the amount of informal classroom visits was a more practical way to gain information about the professional needs of their teaching staffs.

The participants wanted more science staff collaboration and developmental opportunities to improve content specific pedagogical practice, and suggested these be incorporated into professional evaluations. The administrators emphasized the desire to exercise their leadership skills by developing their staff to meet specific needs. The science teachers echoed this desire to improve specific aspects of their teaching. They discussed how professional development would be more helpful to motivate them intrinsically to refine their teaching practice. This was not a new development. Marzano (2012) found that teachers believed that evaluations were important but the bulk of the evaluation should be on professional development and this study had similar findings. Marzano (2012) found that the majority of the survey subjects believed teacher evaluations should be used for both measurement and development, however, development should be prioritized and should be the most important purpose of a teacher’s evaluation.

The participants in this research study echoed the views of the participants in Marzano’s study, which adds to the research on how teachers buy into the evaluative process. This study was different from Marzano’s in that the focus was specifically on secondary science teachers.
and administrators, not teachers in general. Science teachers are unique in that they are content specialists in one of four disciplines (biology, chemistry, Earth science, physics), and according to this study, they desired more discipline specific professional development. The science teachers also expressed their preference to be evaluated by administrators who shared their content specialization. Also, these science teachers in New York State all taught students who took high stakes tests in living environment or chemistry, an important consideration since these assessment measures were part of their APPR composite scores. Perhaps science teachers, as unique as they are, require a different form of evaluation not currently available.

The final key finding involved the use of student test scores as a means to rate teacher effectiveness. The interview data revealed that teachers lacked explicit understanding and largely disagreed with the use of student test scores in their evaluation composite. The data from this study suggested that student growth and achievement should be carefully considered in a science teacher’s evaluation because of the variability among students, lack of seriousness among students taking pre- and post-tests, and the limited reliability and validity of baseline measures. Some of the participants believed that high-stakes state exams did motivate teachers to perform and that teachers should be held accountable for student outcomes to some degree. Others pointed out the use of test scores should only be used to help the science teacher focus on problematic content to better their teaching practice.

The findings gathered from this study were similar to the findings of Jiang, Sporte, and Luppecu’s study on Teacher Perspectives on Evaluation Reform: Chicago’s REACH Students (2015). The researchers determined that the participating Chicago teachers had positive perceptions about the observation process under the evaluation policy and showed apprehension about the addition of student test scores in their evaluation (Jiang et al., 2015). The way science teachers were evaluated in Chicago is different from the way science teachers were evaluated in this study. The three categories for teacher evaluations in Chicago were: 1) 75% for professional practice/observation, 2) 20% for performance tasks, and 3) 5% for individual valued add (Jiang et al, 2015). Chicago and New York have different testing cultures and different graduation requirements. Science teachers in New York were evaluated using a high stakes Regents exams while Chicago science teachers were evaluated with performance tasks. Both studies suggested that policy makers address the divide between positive perceptions of the observation process and the negative views of student test data as a means of evaluation. The authors of the study conducted in Chicago urged additional studies on educators’ perceptions about evaluation reform and the influences upon these perceptions. This study added to the study done in Chicago by examining a context where high stakes exams were a major component of teacher assessment and the teacher perceptions associated with this sort of evaluation.

**Approaches to Science Teacher Evaluation Reform**

**Top-down vs. bottom-up approaches to school accountability reform.** For policy changes to be accepted by the educational community and implemented properly, reform efforts need to be a blend of bottom-up and top-down approaches (Fullan, 1994). Goodlad (1992) observed, “Top-down, politically driven education reform movements are addressed primarily to ‘restrict’ with little to say about educating” (p. 238). Corbett and Wilson’s (1990) study found that mandatory top-down testing requirements in Maryland and Pennsylvania narrowed the curriculum and reduced teacher motivation and collegiality. Fullan (1994) also noted that organizations themselves were not likely to initiate change with the lack of external factors. Beer,
Eisenstat and Spector (1990) showed that top-down and bottom-up approaches to change, when implemented alone, were both flawed. Researchers have concluded that change occurs when centralized mandates and local initiatives unite (Fuhrman, Clune, & Elmore, 1988). Fullan (1994) concluded based on prior research that systems change when individuals and small groups of individuals find commonality both logically and centrally.

A major finding in this study was the APPR system was often not implemented in school districts the way the state intended. There was little teacher "buy in" because of the lack of teacher and administrator input into the policy before the teacher evaluation law was changed. For reform efforts to be successfully implemented, the policy must be clearly outlined for teachers to meet the spirit of the reform. The policy should be practical, providing the teacher with valuable information to improve her teaching practice. Teachers’ and administrators’ input could offer valuable perspectives for policy clarity and practicality in future reform efforts.

Science teacher and administrator perceptions hold the key to successful implementation of future reforms in teacher evaluations. Educator involvement in the design of teacher evaluations will lead to a smoother transition and science teacher agency. Using the information gathered from this study, successful implementation of teacher evaluation reform should correspond with positive teacher perceptions and will focus on intrinsically motivating science teachers to improve instruction. Intrinsically motivating teachers through professional developmental strategies would fulfill one of the goals of Race to the Top – maximizing student learning through improved teacher effectiveness.

**Collective bargaining considerations.** Teacher unions and collective bargaining make statewide uniform methods of science teacher evaluation somewhat challenging. Reform efforts need to take into account the considerable influence that teacher unions have on local teaching contracts. The locally controlled nature of school districts in Suffolk County has created variability in state mandated evaluations as each school district had its own unique teacher contract bargaining unit. Reform efforts must be more general in New York to account for the variability among districts and for the policy to be implemented as intended. The goal of such state policy should focus on teacher development and accountability for student outcomes should be left to the districts to negotiate individually as each district has different needs. A framework for science evaluation generated by educators could be tailored for the needs of individual districts while keeping the goals of the policy consistent.

**Implications**

**Science teachers.** This research study was grounded in the science classroom and the information gathered from this study will impact science teaching in a practice-based manner. This study addressed the concerns of practitioners directly involved in science education. The data support the idea that science teachers are motivated to become better educators and are looking for the proper mechanisms to do so. The results imply that science teachers value the importance of evaluations but would like the focus of the evaluative process to be on professional development, with minimal dependence upon student test scores. The data collected in this study could drastically impact science education at the classroom level as teachers engage in meaningful, practice based, and content specific professional development.

The main goal of APPR was to provide objective teacher evaluation results that would guide professional growth and learning. This goal could be accomplished if the evaluation were geared
towards facilitating that goal. The data collected imply that APPR did not meet that goal. The goal of the science teacher evaluative process should be to improve science education, and teachers need to be consistently motivated to do so. Professional development is an important means for improving the pedagogical content knowledge of science teachers and addressing the dynamic nature of science education.

**Administrators.** The data imply that administrators should serve as science teacher leaders and not merely teacher evaluators. The administrator should determine the needs of the teaching staff and how to fulfill those needs through appropriate professional development. Science administrators expressed the desire to provide their teachers with disciplinary developmental opportunities, however, they were overworked with observations and paperwork. Both science teachers and administrators agreed that conversations about lessons and science education had improved since APPR was implemented and the focus of teacher evaluations should be improved pedagogical practice. How are we going to improve science education? What can we do as a team of educators to improve science instruction? These questions are essential to enhance science teaching and student learning.

**Science teacher educators.** The strength of this study lies with its practical implications for science education within the science classroom and its direct ties to teacher practice. The results of this study imply that professional teachers found the observation rubrics to be useful when discussing aspects of their lessons with their administrators. Pre-service teacher programs could also benefit from this study. These programs could introduce these observation rubrics to pre-service teachers to familiarize them with metrics for lesson evaluations. Highly motivated veteran teachers are calling for more content specific professional development, so it seems logical that pre-service teachers would also require the same training in disciplinary pedagogical content knowledge. Bridging the gap between pre-service science teacher training and professional science teacher development could improve the quality of novice science teachers as they enter the profession and fulfill the goal of improving science learning. One goal of APPR was to ensure that every student received a highly qualified teacher and to provide an evaluation in which the results could guide professional learning. This mirrors the goals of science teacher educators in preparing pre-service teachers to improve practice throughout their careers through reflection and professional growth.

**State policy.** The science teachers in this study all demonstrated a passion for teaching science and having a positive impact on student learning. They found this passion to be the main motivating factor in their careers. Since the goal of this study is to inform policy changes on science teacher evaluation, the focus of the reform should be placed more on the aspects of teaching that will intrinsically motivate teachers to perform. Teachers will be more likely to grow professionally through reform efforts if the emphasis is placed on intrinsic incentives rather than a quantitative effectiveness score. The science teachers and administrators believed that instructional rounds, peer observations, and science educator collaborative communities would achieve that goal.

An essential consideration is how to measure student performance and incorporate this into the evaluation process. The participants cautioned the use of students test scores as a means to evaluate teacher effectiveness and felt the focus of evaluations should be on development rather than measurement. If student test scores are to be used in a teacher’s evaluation, they should be
weighted less when connected to teacher effectiveness; also, variations in student population should be considered. Student test scores should be used formatively as a way to adjust teaching practice in areas in which students have difficulty. Documentation on what the teacher could change to better student understanding of topics would be more beneficial then providing the teacher with an effectiveness score with sometimes arbitrary pre-assessment measures. Student performance is an important part of educational accountability and should be utilized to inform educators about successful and unsuccessful teaching practices. However, this needs to be based upon reliable and valid metrics.

The information gathered from this study suggests strategies for the state to improve science teacher evaluations. Teacher educator experts in professional development and pedagogical content knowledge should be included in rewriting state policy on science teacher evaluations. The recommendations made by the participants is to provide more time when implementing new educational policies, with a pilot phase before systemic implementation. For science teachers to embrace policy changes they must have involvement in policy recommendations. The policy will have a better chance of success if science teachers are clear about the policy, cognizant of useful information the policy provides, find the policy practical to implement, and consider the policy intrinsically motivating.

School district/local policy. This research has laid the foundation not only for impacting state education policy but school district policy. School districts need to devise practical and fair professional development at the local level. Based on the recommendations of science teachers and administrators, the evaluation process should be streamlined with the focus more on teacher and administrator collaboration at little cost to the district. For this study context and similar districts, the number of science teachers within the school districts allows for educational rounds and peer observations to be feasible. Rural districts or small urban schools may lack the number of science teachers needed for this recommendation and they may need to establish inter-district or intra-district collaborative networks. Teachers could observe and collaborate with colleagues throughout these networks, creating subject specific professional learning communities. These learning communities would allow educators to share best practices in the science classroom. Sharing practices with colleagues could rejuvenate seasoned educators who could in turn offer valuable experience to the novice teachers. Veteran science teachers would bring pedagogical content knowledge to the group, a component that was lacking with administrators who did not share the same disciplinary expertise as the individual they were observing.

Administrators could exercise their leadership training in facilitating these collaborative communities, by leading discussions within the group or providing subject specific workshops. The role of an administrator is to be an educational leader, not strictly a subjective observer. The science teachers could keep reflective journals to document their experiences. Since teachers did find the observation process useful, reducing the number of observations done by the administrator would reduce the workload on the administrator and free time for professional development, productive discussions, and informal classroom visits. Educational rounds could be the means to which the collaborative community comes together.

Science teachers felt that the best way to guarantee an effective teacher in every science classroom was through professional development. Overworked administrators lacked the time to execute meaningful professional development, and as a result one of the major goals of Race to the Top and APPR was not fully met. State and local authorities, in conjunction with science teachers and administrators, have the ability to create a meaningful framework for science
teacher evaluations. This framework could focus on the intended goal of APPR, providing objective evaluations that may guide professional learning through a professional development component.

Limitations

The study has several limitations that need to be addressed. Within qualitative research, the researcher acts as both the data collector and the data analyzer, providing the possibility for researcher bias (Miles & Huberman, 1994). The researcher’s background as a science teacher influenced the design of this study and the way the researcher viewed the data. The researcher personally experienced the phenomena of APPR during the teacher evaluation transition period. To bracket her out of the study, the researcher practiced reflexivity, by keeping a self-reflection journal. This technique was used to reduce her bias when analyzing the participants’ experiences with APPR. Although the researcher took steps to minimize her bias, her views could have persisted during data collection and analysis, and could have influenced participant responses during the interviews.

The context where this research study took place was unique compared to the organization of other school districts in the U.S. The research area captured the characteristic features of a much larger school district and had a long tradition of state-mandated test-based accountability. The results from this qualitative study may not be generalizable outside of the state of New York or in rural and small urban school districts and in school districts with different standardized testing cultures.

It is understood that the small number of interviews conducted and the results gathered might not be generalizable beyond the teachers interviewed. The many demands for teachers and administrators inside and outside of school hours limits this study with the amount of time the participants could devote to interviews. This study contained a selected population of teachers; three of the five science teacher participants were members of a selective statewide science teacher collaborative network. These teachers were highly motivated and could attest to the benefits of collaborative teacher networks and the role they play in professional learning. Although the perspectives of all classroom teachers and administrators were not included, their views would likely present more understanding and broader explanations into APPR district policies. A quantitative science teacher perception survey could generate a larger, randomized sample of participants and could make the results of this study more generalizable. Although there are limitations accompanying this study, the results of this study lay the groundwork for future studies on science teacher evaluations.

Systematic Structural Changes to Science Teachers Evaluations

A one size fits all approach to science teacher evaluations is not appropriate for states with large numbers of school districts with varied ranges of socioeconomic diversity. The results generated by the research questions from this study lend themselves to a bigger question, that is, how should science teachers be evaluated?

Structural changes in the format of science teacher evaluation are necessary to accomplish the goals set by Race to the Top and New York State, namely, to provide objective teacher evaluation results that would guide professional growth and learning. Experts in teacher evaluations and pedagogical content knowledge should be included in the revamping of science teacher evaluations. Science teachers and administrators have insightful recommendations to
offer on evaluations, as they are the practitioners of policy, and should be included in the process. More time should be devoted when implementing new science teacher evaluation policy, with a localized pilot phase before systemic implementation.

When it comes to specialized content areas, such as the sciences, educational leadership loses some of its generalizability. Science teachers are content specialists and administrators sometimes do not share the same content expertise of all the science teachers they supervise. The certification of science administrators does not need to be changed or become more specialized, rather teacher leaders and peer learning communities could lesson the burden on science administrators, reducing the amount of time dedicated to formal observations and write-ups. The administrator’s role could be adjusted to become a facilitator for peer observations, teacher collaborative communities, and educational rounds. The administrator would change his role from strictly an evaluator to an organizer, seeking out meaningful professional development to fulfill the needs of faculty.

The results of this study indicate that the professional development of science teachers should be the focus of teacher evaluations. The participants found accountability to be an important component of evaluations, yet these evaluations may not need to be conducted every year. Composite scores calculated over a longer period of time could provide a better representation of the teacher’s quality. Developing and documenting individual professional learning plans could be a way to evaluate science teachers in the future. Rather than relying on flawed metrics for student growth and achievement, other evidence should be considered. Summative assessments completed by school leaders or multiple assessments within the school year could provide the information needed to demonstrate student growth within the science content. Table 13 provides a list of recommendations based upon the results of this study.
Table 13
Summary of Recommendations

- Teacher educator experts in professional development and pedagogical content knowledge should be included in rewriting state policy on science teacher evaluations.
- The cautious nature of teacher perceptions associated with student test scores and the positive perceptions associated with the observation process indicate that policy makers should address this dichotomy in future evaluation reform efforts.
- The one-size fits all model of teacher evaluation is not appropriate in states with a large number of diverse school districts. A general, statewide framework for teacher evaluations could provide school districts the means to make decisions about evaluating their teachers within that framework.
- Developing science teachers should be the goal of teacher evaluations.
- Meaningful practice based, content specific professional development is the key to motivating science teachers intrinsically to change their practices.
- The rise of content specific teacher leaders could provide the disciplinary expertise required for science teacher development and may also alleviate some of the burden on administrators.
- Current administrators should serve as facilitators for science teacher networks and peer observations.
- Administrators should seek out professionals to run workshops based on the needs of their science teaching staff.
- Teacher composite evaluation scores should be an aggregate of more than one year of data; perhaps three or more years of data would provide a better representation of teacher quality.
- Peer observations and science teacher collaborative networks, both within school districts and across school districts, can drive professional learning.
- Science teacher professional learning plans could be used as a part an evaluation.

Future Research

The findings from this study indicate possible areas for future research. Additional studies on the impact of instructional rounds on science education could support the implementation of this recommendation. There has been no research on the impacts of educational rounds on science teachers and science learning. The influence of educational rounds and science teacher motivation could also be studied. Before implementing educational rounds into policy, studies would need to be conducted on implementing rounds among science teachers and the effects it had on motivation, science teaching practice, and student outcomes. A future study relating to administrators’ roles in the practice of education rounds and the impact of student growth and achievement would be insightful for scaling evaluation reforms.

Standard teacher observation rubrics are a one size fits all solution to evaluate a teacher’s effectiveness. The rubrics used to evaluate science teachers in the participating districts did not include a pedagogical content knowledge component. Shulman (1986) identified pedagogical content knowledge as an important knowledge base for teaching, since it encompasses the teacher’s understanding of how students learn specific subject matter. Pedagogical content knowledge has been found to be a complex process that is extremely specific to context, situation, and person. Because of its highly specific nature, it has to be closely aligned to teachers’ professional practice in a personal way. This can be achieved by providing teachers with input, discussion of instructional strategies, self-reflection, and group reflection (Van Driel & Berry, 2011). Content specific professional development can improve pedagogical content knowledge in more meaningful ways and lead to greater teacher effectiveness and student learning.
Science teachers are content specialists, most having a four-year degree in a science, and they require more detailed evaluation rubrics to reflect the diversity of subjects taught (biology, chemistry, physics, Earth science). However, few are in circulation. Some of the participants recommended more content specific strategies to help improve their educational practice and content specific rubrics could provide these strategies. These teachers were motivated to improve and were searching for the guidance to do so.

This study only elicited the suggestions and recommendations of the participating science teachers and administrators. A future study reflecting different techniques to evaluate science teachers would be beneficial in creating a framework for science teacher evaluation policy. Do science teachers need to be evaluated every year the same way? Marzano (2012) suggested that the Rapid Assessment of Teacher Effectiveness (RATE) could be used as a scoring rubric and development scale. Science teachers could be evaluated informally using the RATE rubric to efficiently determine teacher competence (Strong, 2011). Could the RATE rubric be used to evaluate informal observations? The improvement of science teachers is more complex and requires continuous content specific professional assessment, which cannot be determined in one classroom observation. A study relating the use of the RATE rubric to rapidly evaluate science teachers and the impacts on science education could also provide useful information to the educational community.

Conclusions

Race to the Top influenced the way science teachers were evaluated in New York. The goal of the legislation was to guarantee that there was an effective teacher in every classroom and this became the goal of New York teacher evaluations. New York changed the way teachers and school leaders were evaluated and APPR was designed to encourage professional growth and improve instructional practices. According to this study, APPR has had questionable success in achieving these goals for science teachers in Suffolk County.

The goal of science teacher evaluations should be to provide the framework for developing excellent science educators. That framework could be used at the local district level to fulfill the professional needs of the teachers within that district. This study has established some important considerations when designing and implementing science teacher evaluations. Prior research is limited regarding teacher insights in evaluation reform. The research has shown that teachers and administrators have the greatest impact on student performance because they are the closest to the learning process. These motivated individuals have the means and understanding at the classroom level to make positive changes in science education for all students.

Several important points were learned as a result of this study: 1) science teachers and administrators value the importance of teacher evaluations, 2) conversations between administrators and teachers have improved as a result of APPR’s implementation of observation rubrics, 3) the condensed time-line to implement APPR policy may have contributed to teacher resistance because of lack of policy clarity, 4) science teachers and administrators found APPR did not provide them with practical and useful information to improve science teacher practice with the use of student test scores, and 5) science educators believed professional development should be the main focus of teacher evaluations.

The insights gathered from this study add to the literature regarding science teacher and administrator perspectives of evaluation policy, an area in the literature that is lacking in recent educational accountability movements. As teacher evaluation has gained attention across the
country, adding teacher and administrator perceptions to the literature provides the foundation for future studies in this area. This research calls for future studies regarding: 1) what we know about professional development; 2) how professional development could be organized into teacher evaluations; 3) content specific professional development and its impacts on science education; 3) the effects of professional development on teacher motivation; and 4) the impacts of science teacher collaborative networks, peer review, and educational rounds on science teaching and learning.

Overall, this study found positive potential in the evaluations of teachers in New York under the APPR system. Teachers and administrators reported that conversations as a result of the implementation of lesson observation evaluation rubrics had improved. The addition of these rubrics gave science teachers expectations of evaluation metrics and gave the administrators the means to measure lesson quality. However, the teachers and administrators were concerned about the use of student test scores to measure teacher competence. This trepidation may have been due to the lack of clarity regarding how these scores contributed to the composite score or the lack of practicality that the student scores provided. The dichotomy between APPR’s limitations with the use of student test scores and science educators’ positive perceptions of the conversations generated during the observation process provides valuable information to policy makers. The cautious nature of teacher perceptions associated with student test scores and the positive perceptions associated with the observation process indicates that policy makers should address this dichotomy in future evaluation reform efforts.

Future reform efforts regarding science teacher evaluations should include educator input, explicit and valid metrics, and possess practicality for science teachers and administrators. The evaluation policy should emphasize disciplinary professional development and teacher professionalism to motivate teachers intrinsically and foster continuing education. The contributions of this research study will guide future studies in science teacher evaluation reform and teacher/administrator perceptions of educational evaluation policy implementation.
References


New York State Education Department. (2015b). *Guidance on New York State’s annual professional performance review for teachers and principals to implement educational law 3012-c and the commissioners regulations*. Albany, NY: NYSED. Retrieved from...


Appendix A: Science Teacher Semi-Structured Interview Protocol

1. Comment on your experiences with the new observation process. (Danielson Rubric)
2. Did your school district explain the new process in detail?
3. Talk about your perceptions of the local achievement measure.
4. Talk about your perceptions of the student learning objectives.
5. What is your general impression of the overall system?
6. Do you feel APPR fosters a culture of continuous professional growth?
7. Talk about what motivates you in your career.
8. Which science content area and level do you teach?
9. What levels of science courses does your district offer (AP, IB, honors level, Regents Level)?
10. Do you feel as though you are an effective science teacher? Why or why not?
11. What qualities do you think make up a great teacher?
12. Do you participate in any activities outside the school day for the school district you work in?
13. Do you think student test scores should be used to evaluate teacher quality?
14. Describe your feelings about the new Annual Professional Performance Reviews.
   a. Do you think the current system is a fair way to evaluate a teacher?
15. What would be, in your opinion, the ideal way to evaluate teachers in Suffolk County?
16. Have you changed anything about the way you teach because of the new APPR system of evaluation?
   a. Do you do more Regents Test Prep?
17. Would you be opposed to a student/parent survey as a part of your evaluation?
18. How do you feel about the addition of a teacher self evaluation as a part of your evaluation?
Appendix B: Administrator Semi-Structured Interview Protocol

1. Tell me about your background.
2. What certifications do you hold?
3. Tell me about what motivates you in your career.
4. How are you evaluated?
5. Tell me about your understanding of how you are to evaluate science teachers in your district.
6. How many teachers are you responsible for?
7. Comment on your experiences with the new observation process.
8. Talk about your perceptions of the local achievement measure.
9. Talk about your perceptions of the student learning objectives.
10. Do you think student test scores should be used to evaluate teacher quality?
11. Tell me about your general impression of the overall evaluation system?
12. Do you think teachers have changed anything about the way you teach because of the new APPR system of evaluation?
13. Do you feel APPR fosters a culture of continuous professional growth?
14. Tell me about the qualities that you think make up a “great” teacher.
15. What are your views on tenure?
16. Describe your feelings about the new Annual Professional Performance Reviews.
17. What would be, in your opinion, the ideal way to evaluate teachers in Suffolk County?
Dear ________________,

I am a doctoral student of Science Education at Stony Brook University, and I am planning on conducting a research study to examine teacher perceptions of the current Annual Professional Performance Reviews (APPR) for Regents science teachers in Suffolk County New York. I hope to provide insights to impact policy regarding teaching evaluations and to reform teacher evaluation practices. Potential participants will include current secondary science teachers with varied ranges of experience and administrators responsible for science teacher evaluations in a public school district in Suffolk County New York.

I am soliciting potential participants by contacting current secondary science teachers via email and asking them to participate in semi-structured interviews. Participants will be required to sign informed consent as a condition for being part of the study. The identities of all participants and their schools and cities will be kept confidential. All interview transcripts will be destroyed three years after the study is completed.

The results of my research will be available to you once my work is completed. I have secured approval for research with human subjects from the Institutional Review Board at Stony Brook University (attached).

Please contact me at Jessica.Mintz@stonybrook.edu or 631-484-2617 (cell) should you have any questions.

Thank you very much for your consideration.

Sincerely,

Jessica A. Mintz

Center for Science/Mathematics Education
Appendix D: Consent Form

Project Title: Science Teacher & Administrator Perceptions of the Annual Professional Performance Review in Suffolk County New York.
Principal Investigator: Angela M. Kelly, Ph.D.
Co-Investigators: Jessica Mintz, MAT
Department: Center for Science & Mathematics Education

You are being asked to be a volunteer in a research study.

PURPOSE

The purpose of this study is to examine teacher and administrator perceptions of New York State teacher evaluation for science teachers, the Annual Professional Performance Review (APPR). The study also plans to investigate the unintended consequences the teachers and administrators report using the new APPR system. The main goal of the study is to impact policy regarding teaching evaluations and to reform teacher evaluation practices using fair, consistent and research based methods. You are eligible for this study because you are an administrative employee or high school science teacher in a Suffolk County school district. There will be approximately 20 participants in this research study.

PROCEDURES

If you decide to be in this study, your part will involve:

- A questionnaire regarding your demographic information, as well as general teaching information. This may require approximately 30 minutes.
- Three to four 30-45 minute interviews regarding your thoughts on science teacher evaluations in New York.
- Audiotaping of the interview, or if you prefer, the researchers will take field notes during the interview.

RISKS / DISCOMFORTS

There are no foreseeable risks or discomforts associated with your participation in this study.

BENEFITS
There is no direct benefit expected as a result of you being in this study. There may be indirect benefits in terms of research data that will improve policies related to the evaluation of teacher quality.

CONFIDENTIALITY

Protecting Your Privacy in this Study

We will take steps to help make sure that all the information we get about you is kept private. Your name will not be used; we will use a code instead. All the study data that we get from you will be kept locked up. The code will be locked up too. If any papers and talks are given about this research, your name and the name of the school district will not be used. All audiotapes of interviews will be deleted after transcription.

We want to make sure that this study is being done correctly and that your rights and welfare are being protected. For this reason, we will share the data we get from you in this study with the study team, Stony Brook Universities Committee on Research involving Human Subjects, applicable Institutional officials, and certain federal offices. However, if you tell us you are going to hurt yourself, hurt someone else, or if we believe the safety of a child is at risk, we will have to report this. In a lawsuit, a judge can make us give him the information we collected about you.

COSTS TO YOU

There are no costs associated with participation in this study.

YOUR RIGHTS AS A RESEARCH SUBJECT

- Your participation in this study is voluntary. You do not have to be in this study if you don't want to be.

- You have the right to change your mind and leave the study at any time without giving any reason, and without penalty.

- Any new information that may make you change your mind about being in this study will be given to you.

- You will get a copy of this consent form to keep.

- You do not lose any of your legal rights by signing this consent form.

QUESTIONS ABOUT THE STUDY OR YOUR RIGHTS AS A RESEARCH SUBJECT

- If you have any questions, concerns, or complaints about the study, you may contact Jessica Mintz, (631) 484-2617, OR Jessica.Mintz@stonybrook.edu or Dr. Angela Kelly, (631) 632-9750, OR angela.kelly@stonybrook.edu.

- If you have any questions about your rights as a research subject or if you would like to obtain information or offer input, you may contact Ms. Judy Matuk, Committee on
Research Involving Human Subjects, (631) 632-9036, OR by e-mail, judy.matuk@stonybrook.edu.

If you sign below, it means that you have read (or have had read to you) the information given in this consent form, and you would like to be a volunteer in this study.

Subject Name (Printed)

Subject Signature   Date

Name of Person Obtaining Consent (Printed)

Signature of Person Obtaining Consent   Date