Examining the Development of Beginning Middle School Math Teachers' Practices and their Relationship with the Teachers' Effectiveness

By

Laura Neergaard Booker

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Leadership and Policy Studies

May, 2014

Nashville, Tennessee

Approved:

Thomas M. Smith, Ph.D.

Ellen B. Goldring, Ph.D.

Mimi Engel, Ph.D.

Marcy Singer-Gabella, Ph.D.

To beginning teachers working tirelessly to meet their students’ needs

ACKNOWLEDGEMENTS

This dissertation marks the culmination of my time as an official student. However, if there is one thing I have truly learned, it is that I still have a lot to learn. I may be done with school, but I will never be done learning. Still, getting to this point required a lot of support from a variety of people, and I would like to recognize them for helping me to reach this goal.

First, I want to thank Teach For America and the Leland Public School District for providing an opportunity for me to be a teacher. That experience opened by eyes to the struggles of beginning teachers, sparking my desire to figure out how to better support novice teachers in developing their skills.

Next, my dissertation chair and committee have provided valuable feedback to help me focus all the ideas running through my head. This process has been challenging in ways I never anticipated, and I appreciate their understanding and consistent high expectations. I also want to thank the AIM team for their efforts designing surveys and interview protocols, collecting data, and watching hours of video of teachers' instructional practice. Thanks especially go to Tom Smith, Laura Desimone, Marisa Cannata, Katherine Taylor Haynes, Eric Hochberg and Mary Batiwalla. I also want to thank the beginning teachers who opened their classrooms to us.

Many others in the Leadership, Policy, and Organizations community played a role in helping me develop and hone my research skills. I am grateful for their guidance. I especially want to thank those colleagues who assured me that it was perfectly reasonable to take a job that I felt was most connected to teachers and students. I would also like to acknowledge my current colleagues at the Tennessee Department of Education for their kindness and understanding of my endeavor to finish my dissertation while working a fulltime job.

I always have told prospective students that one of the best things about my doctoral program was the other students. Among my many wonderful peers, Courtney Preston was my partner through this process. Courtney, you made this journey so much better.

Finally, I want to thank my close friends and family who were persistent in their inquiries about my dissertation. You held me accountable, and I am grateful for your interest and encouragement. I especially want to acknowledge my mother who instilled in me the importance of education and a commitment to our public schools. She got me through the last stage of this process by serving as my copy editor. I occasionally doubted myself, but she never did. I also appreciate the patience and support of my husband who entered into this process midway. Mike, I appreciate you understanding how important it was for me to reach the finish line and your gentle nudges to help me get there.

TABLE OF CONTENTS

Page

ACKNOWLEDGEMENTS iii

LIST OF TABLES vi

LIST OF FIGURES viii

Chapter

I. INTRODUCTION 1

II. LITERATURE REVIEW 7

Teaching Practices 8

Development of Beginning Teachers’ Practices 13

Relationship of Teaching Practices to Student Achievement 17

Gaps in Previous Research 33

III. METHODS 35

Participants 35

Measures 37

Analysis Methods 56

IV. RESULTS 67

Development of Beginning Teachers’ Practices 67

Relationship between Classroom Environment and Instructional Teaching Practices . . 74

Relationship between Teachers’ Practices and Teacher Effectiveness 85

V. DISCUSSION 100

Considerations 100

Review of Findings 102

Policy Implications 107

Future Research 112

VI. APPENDIX 116

VII. REFERENCES 124

LIST OF TABLES

Table 1. Relationship between Student Achievement and Measures of Teaching Practice from MET Project 31

Table 2. Characteristics of Participating School Districts 36

Table 3. Teacher Background Characteristics from AIM and 2007-08 SASS 37

Table 4. Number of Teachers Recorded at Each Time Period 38

Table 5. Overview of the CLASS-S Domains, Dimensions, and Indicators 41

Table 6. CLASS-S General Scoring Guidelines 42

Table 7. CLASS-S Quality of Feedback Rubric 42

Table 8. Overview of the Math-specific IQA Rubrics and Guiding Questions 47

Table 9. Math-specific IQA Scoring Guidelines 47

Table 10. Alignment between Math-specific IQA Rubrics and CLASS-S Dimensions 52

Table 11. CLASS-S and IQA Mean, Minimums, and Maximums over Time 72

Table 12. Growth Coefficients Indicating Monthly Improvement in Teaching Practice 73

Table 13. Correlations between Overall IQA Scores and CLASS-S Scores over Time 76

Table 14. Correlations between IQA Rubric Scores and CLASS-S Instructional Support Scores over Time 76

Table 15. Correlations between CLASS-S Domain Scores over Time 76

Table 16. Predicted Probability of “High” Instructional Practices Given Classroom Environment Practice Scores from the Same Observation Period (HLM) 82

Table 17. Predicted Probability of “High” Instructional Practices Given Classroom Environment Practice Scores from the Same Observation Period (Fixed Effects) 83

Table 18. Predicted Probability of “High” Instructional Practices for Prior Classroom Environment Practices from the Prior Observation Period 84

Table 19. Correlations between Teachers’ Value-added Scores and Practice Ratings 85

Table 20. Correlations between Teachers’ Value-added Scores and Practice Ratings for Stayers 86

Table 21. Correlations between Change in Value-added Scores and Change in Practice Ratings 87

Table 22. Correlations between Value-added Scores and Practices Ratings at the Dimension and Rubric Level 90

Table 23. Teaching Practice Ratings by Value-added Score Quartiles 91

Table 24. Predicted Association between Value-Added Scores and Practice Ratings 92

Table 25. Predicted Association between Value-Added Ratings and Practice Rating, Controlling for Other Practice Ratings 94

Table 26. Predicted Association between Value-Added Scores and Practice Ratings, Controlling for Time-Invariant Teacher Characteristics (Fixed Effects) 96

Table 27. Predicted Association between Value-Added Scores and Practice Ratings, Controlling for Other Practice Ratings and Time-Invariant Teacher Characteristics (Fixed Effects) 97

Table 28. Predicted Association between Change in Value-Added Scores and Change in Practice Ratings 98

LIST OF FIGURES

Figure 1. Teachers Rated as Low, Mid, or High on their Fall Year 1 Teaching Practices 68

Figure 2. CLASS-S Scores over Time 70

Figure 3. Math-specific IQA Scores over Time 70

Figure 4. CLASS-S Domain Means by Rigorous Math Implementation over Time 78

Figure 5. CLASS-S Domain Means by Rigorous Math Implementation over Time 78

Figure 6. Scatterplot of IQA Task Implementation and CLASS-S Classroom Organization Scores 80

Figure 7. Scatterplot of IQA Task Implementation and CLASS-S Emotional Support Scores 81

CHAPTER I

INTRODUCTION

Research consistently finds that teachers' effectiveness improves during the first few years on the job (Clotfelter, Ladd, & Vigdor, 2007; Harris & Sass, 2007; Henry, Bastian, & Fortner, 2011; Kane, Rockoff, & Staiger, 2006). These increases are typically attributed to improvements in the novice teachers' teaching practices. Yet, we know little about which practices new teachers are most likely to improve during their first years of teaching, or, more specifically, about the process through which this improvement takes place.

While no consensus exists on what effective teaching consists of, research has identified several practices that are associated with increased student learning (Grossman et al., 2010; Hill, Kapitula, & Umland, 2011; Holtzapple 2003; Kane, Taylor, Tyler, & Wooten, 2010; Kimball, White, Milanowski, & Borman, 2004; Milanowski 2004). These practices are often grouped under two constructs: classroom environment and instruction (Danielson, 1996; Doyle, 1986). Classroom environment typically includes: (a) practices to build classroom culture and foster a positive learning environment and (b) practices to manage behavior and time (Hamre, Pianta, Mashburn, & Downer, 2008). Instructional practices also tend to fall into two groups: cross-subject and content-specific (Grossman et al., 2010; Stodolsky, 1988). Cross-subject instructional practices are applicable across subject areas, whereas content-specific instructional practices are specific to a particular subject area.

Studies of quality of teaching practices usually focus either on a particular type of teaching practice (e.g., Matsumura et al., 2006) or an average rating of teaching quality across a variety of practices (e.g., Holtzaple, 2003; Schacter & Thum, 2004). Few rigorous studies have differentiated types of practices, making it difficult to determine precisely which practices are more likely to produce learning gains. Research examining both cross-subject and content-specific instructional practices is especially rare. Thus, uncertainty remains around what combination of teaching practices are most likely to influence student achievement.

Assessing teaching quality across a broad range of practices is important for understanding beginning teachers' proficiency at a variety of teaching practices when they first enter the profession, and knowing where beginning teachers start is also necessary for studying the development of their practices over time. Considering multiple types of teaching practices can provide insight into which practices new teachers quickly master, which practices they struggle to improve, and whether development of some practices is related to growth in others. For example, qualitative research on pre-service and beginning teachers has suggested that attaining proficiency in classroom management practices is often necessary for these teachers to reach higher levels of instructional quality (Kagan, 1992). However, this notion has not been explored empirically. Data from ratings of various teaching practices can be used to study trends in the relationship among teaching practices, such as whether mastery of classroom environment practices is associated with attaining proficiency in instructional practices.

Examining a wide range teaching practices for the same group of teachers can also help us understand what makes some beginning teachers more effective than others. While a growing body of evidence indicates the importance of teachers for producing student learning (Aaronson, Barrow, & Sanders, 2007; Nye, Konstantopoulos, & Hedges, 2004; Rivkin, Hanushek, & Kain, 2005; Rockoff, 2004), we are less clear about what makes one teacher more effective than another. Differences in teachers' contributions to student learning are likely to arise from differences in their teaching practices. Thus, to better understand why some teachers are more successful than others, it is critical to examine the relationship between the ratings of a variety of teaching practices and measures of student learning.

The Gates Foundation’s Measures of Effective Teaching (MET) has taken steps toward a better understanding of which teaching practices may be more important for increasing student learning by studying the relationship of multiple measures of teaching effectiveness. Findings indicated that ratings from some content-specific observation measures were more strongly associated with measures of teacher effectiveness than ratings from cross-subject instruments (Kane & Staiger, 2012). While the information is helpful for identifying the strength of relationships between different types of teacher evaluation measures, the focus of the MET project was on comparing across a variety of observation measures rather than studying the relationship of specific practices to each other and to their relationship with effectiveness at increasing student learning. Thus like most other studies of teaching practices, the MET project does not provide information on the relationship of classroom environment practices and cross-subject and subject-specific instructional practices to teacher effectiveness or information about change in teaching practices over time.

This study uses observational and student achievement data to explore the development of beginning mathematics teachers’ practices and their relationship to teacher effectiveness at increasing student achievement using data from a longitudinal study of beginning middle school mathematics teachers’ induction and mentoring experiences. The Classroom Assessment Scoring System - Secondary (CLASS-S) was used to measure cross subject environmental and instructional practices and the Instructional Quality Assessment (IQA) was used to measure math-specific instructional practices. The study addresses the following research questions:

- What are the initial levels of beginning middle school math teachers’ classroom environment and instructional practices?

- To what extent do beginning teachers improve on various aspects of their classroom environment and instructional practices during their first three years of teaching?
- What is the relationship between teachers' environment and instructional practices?
    - Are beginning teachers who demonstrate strength in their classroom environment practices also strong in their cross-subject and content-specific instructional practices, or are teachers strong in some areas and weak in others?
    - Do classroom environment practices such as behavior and time management need to reach a particular level of competency before teachers can improve cross-subject and content-specific instructional practices?
- To what extent are classroom environment and instructional teaching practices associated with beginning teachers' effectiveness at increasing student learning?

Understanding the development of teaching practices among beginning teachers can guide teacher training and professional development. For example, if teachers tend to show improvement in cross-subject classroom environment and instructional practices, but struggle to improve their subject-specific instructional practices, supports can be targeted to help teachers with these challenges. Moreover, knowing if it is common for new teachers who are strong in classroom environment practices to also be strong in cross-subject and math-specific instructional practices or if it is likely for new teachers to being strong in some areas and weak in others can also inform our thinking about the supports provided to new teachers. Awareness of the variety of patterns to look for in the practices of the beginning teachers can be used to better design supports. In addition, if high quality classroom management practices are identified as a precursor to rigorous instructional quality then supports can be focused on those teachers struggling to reach classroom management proficiency.

Information on which practices may lead beginning teachers to be more effective can support district and school leaders in focusing their policies and procedures on increasing student achievement. For example, when hiring new teachers, school and district administrators may request applicants for teaching positions to model lessons. Knowledge that certain practices would lead to greater learning gains can inform what administrators making hiring decisions look for during these demonstrations. Once hired, beginning teachers are often targeted for supports like mentoring and induction. Identifying whether certain instructional practices of beginning teachers are related to student gains can help improve the focus of those supports. Finally, school leaders can evaluate proficiency in the teaching practices linked to student gains when determining whether a new teacher should receive job renewal or be granted tenure.

Results of this study may also be useful for guiding what teaching practices are included on classroom observation measures used for teacher evaluation. Feedback from evaluations of teacher practices can facilitate critical reflection that leads to improvement (Taylor & Tyler, 2011). Better understanding of the development of beginning teachers' practices can help discern what practices should be included on evaluation measures so that strengths, weaknesses, and improvement can be measured and used to provide constructive feedback. Evaluations should be assessing teacher practices that research has found to be related to student achievement, and knowing which instructional practices are related to student learning gains can help inform teacher evaluation programs.

The next section reviews the literature on the constructs of classroom environment and instructional teaching practices, the development of beginning teachers' practices in these two areas, the relationship between different teaching practices, and the relationship of those practices to teacher effectiveness. Following this, I describe the participants, measures, and analysis methods used to address the research questions. Then, I present the results of the analysis. Finally, I conclude with a discussion of the findings.

CHAPTER II

LITERATURE REVIEW

Studying teaching practices is challenging due to the multidimensional and complex nature of the work of teaching (Regan, Case, Case, & Freiberg, 1993). Teaching practices are intertwined and separating them into distinct categories is difficult. Researchers have used a variety of conceptual frameworks regarding the work of teaching (Lampert, 2010). This study focuses on in-class teaching practices, using Goe's (2007) definition of teacher practices as "what they [teachers] actually do in the classroom with their students" (p. 10). Given this definition, the practices examined in this study are largely defined by teacher actions, as well as by their interactions with students. In-class practices are also those that are visible to an observer, excluding actions like lesson planning prior to class (Grossman et al., 2010).

For this study, I relied on conceptual frameworks of teaching practices to establish the constructs of the in-class teaching practices analyzed. Frameworks of in-class teaching practices typically identify two primary constructs, classroom environment practices and instructional practices, under which most in-class practices can be grouped (e.g., Danielson, 19966; Kane et al., 2010). In this chapter, I define these two constructs of teaching practices and establish the importance of specific practices within these constructs for improving student outcomes. A report on previous research on the development of beginning teachers' classroom environment and instructional practices follows. Then, I discuss findings from prior research that have analyzed the relationship between classroom environment and instructional teaching practices and student learning. This chapter concludes with a discussion of the gaps in existing research that will be addressed by this study.

Teaching Practices

Classroom environment teaching practices. Classroom environment teaching practices "set the stage for all learning" (Danielson, 2007, p. 28) and can be divided into two groups: (a) practices to build a positive classroom culture and (b) practices to manage students' behavior and classroom processes.

Cultural and social/emotional practices. A positive classroom culture includes creating an environment of respect and rapport between the teacher and students as well as amongst students and their peers, establishing a culture for learning, and responding to students' social/emotional needs and developmental levels (Danielson, 2007; Pianta, Hamre, Mintz, 2011). A classroom with a positive culture includes warm, respectful relationships between students and teachers and between peers. In classrooms with a positive culture, teacher communications demonstrate enthusiasm for learning, are encouraging, and lack negativity (Pianta et al., 2011). Social/emotional practices include monitoring students, noticing when they need extra support, then responding to their social/emotional needs. Practices demonstrating regard for students' social/emotional needs support student ideas and provide opportunities for leadership and choice (Pianta et al., 2011).

In classrooms where students feel safe and supported, students are more likely to be motivated and engaged, and these characteristics are related to educational achievement and attainment (Klem & Connell, 2004; Ryan & Patrick, 2001). Students' social and emotional functioning and relationships with their teachers can also influence their academic achievement and educational attainment (Gilman & Anderman, 2006; National Center for Education Statistics, 2002; Wang, Haertel & Walberg, 1997). For example, Crosnoe, Kirkpatrick-Johnson, and Elder (2004) found that stronger student-teacher relationships were associated with higher academic achievement and with a lower likelihood of disciplinary problems for secondary students. Conversely, when teachers lack the ability to effectively manage students' social and emotional needs, their students demonstrate lower levels of on-task behavior and performance (Marzano, Marzano, & Pickering, 2003).

Management practices. The second group of classroom environment teaching practices includes managing student behavior, classroom procedures, instructional formats, and materials (Emmer & Stough, 2001). Behavior management practices focus on the teacher's use of effective methods to encourage desirable behavior as well as to prevent and redirect misbehavior. Productivity practices describe how well the teacher manages time and routines so that instructional time is maximized. Practices to manage student engagement with learning materials and instructional formats involve organization; therefore they are grouped with management practices. Instructional formats are methods and means teachers use to deliver information.

Classroom management practices have long been considered an essential component of a teachers' repertoire of skills. The underlying reasoning for the importance of management practices is that student learning is better able to occur when student behavior meets expectations and students are engaged (Pianta, LaParo, & Hamre, 2004). Research has found that student learning increases more in classrooms with positive behavior management and time management (Brophy & Good, 1986; Coker, Medley, & Soar, 1980; Good & Grouws, 1977; Soar & Soar, 1979). In a study on elementary mathematics instruction, Good & Grouws (1977) found that teachers whose classes had greater achievement gains tended to have better management skills, evidenced by spending less time in transitions and dealing with discipline problems. In a review of studies examining teacher behavior and student achievement, Brophy and Good (1986) concluded that the positive relationship between amount of time spent on instruction and student achievement was the most consistently replicated finding. Research has also shown benefits of providing teachers with classroom management training on observed teacher management practices and student engagement (Evertson, Emmer, Sanford, & Clements, 1983; Piwowar, Thiel, Ophardt, 2013).

Instructional teaching practices. While classroom environment practices “set the stage” for learning, instructional practices support student learning of concepts and skills and ensure students are able to apply the knowledge they have gained. Danielson (2007) refers to the instruction components as the “actual engagement of students in content” (p. 29). For the purposes of this study, I classify instructional practices as either cross-subject or content-specific (Grossman et al., 2010; Stodolsky, 1988). In the research literature, cross-subject instructional practices are practices that can be applied across content areas, whereas content-specific practices are conceptualized differently across subject area.

Cross-subject instructional practices. Cross-subject instructional strategies focus on approaches used to help students understand the academic material; questioning and discussion techniques to facilitate students’ understanding and use of higher-order thinking; and ongoing assessment in instruction (Danielson, 2007; Pianta et al., 2011). Higher order thinking is thinking that is critical, logical, reflective, metacognitive, or creative (King, Goodson, & Rohani, 1998). These teaching practices manifest through a teacher’s delivery of material, selection of classroom assignments, and conversations with students.

Several studies have established relationships between specific cross-subject instructional practices and student learning. Cross-subject instructional practices include presentation of material, tasks and activity selection, questioning, feedback, and assessment strategies. Several studies corroborate that effective teachers present new material in increments with attention to students' prior knowledge (Bransford, Brown, Cocking, 2000; Rosenshine, 1995; Marzano, 2004). Research suggests that more rigorous classroom assignments that require higher-level thinking strategies are associated with higher student performance (Newmann, Lopez, and Bryk, 1998). Teachers who assign hands-on learning activities tend to have students who show greater achievement gains (Wenglinksy, 2002). Evidence also shows that certain types of feedback are more successful at increasing student learning (Hattie & Timperley, 2007). For example, studies validate that feedback is more effective when it is immediate, designed to increase interest and effort, and promote higher-order thinking (Good & Brophy, 2008; Kulik & Kulik, 1988).

Content-specific instructional practices. While research has established several practices that are linked to increased student outcomes across subject areas, other research suggests that effective teaching practices differ between subject areas (Evertson, Anderson, Anderson, & Brophy, 1980; Graeber, Newton, & Chambliss, 2012; Grossman & Stodolsky, 1995; McDonald & Elias, 1976; Stodolsky & Grossman, 1995). Content-specific and cross-subject instructional practices often overlap in their broader goals, but content-specific practices tend to be defined more explicitly than cross-subject practices. For example, while assigning tasks and implementing instruction to elicit higher-order thinking is a teaching practice that is valued across content areas, specific practices that encourage higher-order thinking may differ between subjects (Newmann, Lopez, and Bryk, 1998). Current priorities evident in recent mathematics instructional reform include the use of mathematical tasks with multiple solution strategies for producing higher-order thinking (Cohen & Ball, 1990; Hiebert, 2003; Hill et al., 2011; NCTM, 2000). Teaching practices to facilitate higher-order thinking when teaching reading comprehension include assigning quality texts for students to read, engaging students in the deeper meaning of the texts, and providing tasks that enable students to apply higher-level thinking skills (Snow, 2002).

Research on teachers' content-specific instructional practices finds that both content knowledge and pedagogical content knowledge are critical for effective teaching (Ball, Thames, & Phelps, 2008; Darling-Hammond, 2000; Hill, Rowan, & Ball, 2005; Shulman, 1986). Content knowledge is an understanding of a particular subject area, and pedagogical content knowledge is knowledge of how to teach that subject-matter (Ball, 1990; Borko & Putnam 1996; Cohen, McLaughlin, & Talbert, 1993; National Commission on Teaching and America's Future, 1996; Shulman, 1987). Content knowledge and pedagogical content knowledge are expected to manifest in teachers' content-specific instructional practices (Bell, Gitomer, McCaffrey, Hamre, Pianta, & Qi, 2012).

Few studies have captured differences between effective teaching practices across subject areas. In a study of junior high mathematics and English classes, Evertson at al. (1980) found a positive relationship between teachers asking questions that required explanations in the mathematics classes, but not in the English classes. The authors also reported differences in student ratings of "good teaching" for mathematics and English teachers. Academically demanding mathematics teachers tended to be rated positively by students, whereas academically demanding English teachers tended to be rated negatively. Teachers of different subjects also express differing views of instructional practices. Grossman and Stodolsky (1995) reported findings from surveys and interviews with high schools teachers. They found that mathematics teachers in the study felt that their instructional practice was heavily influenced by the sequential nature of the subject, whereas, the English teachers felt they had more flexibility in deciding what content to cover and how to teach it.

Until recently, few subject-specific observation measures of instructional practices existed (Kennedy, 2010). In a pilot study of a mathematics-specific tool, the Instructional Quality Assessment, Matsumura et al. (2006) found a significantly positive relationship between students' achievement on mathematics tests and their teachers' assigning and implementing more rigorous tasks and leading student discussion of mathematics concepts. Another mathematics-specific tool, the mathematical quality of instruction (MQI) found evidence that teacher effectiveness scores were correlated with instructional practice ratings based on the richness of the mathematics and the absence of errors and imprecision, as well as explicitness and thoroughness in presentation of content (Hill, Kapitula, & Umland, 2010).

In the above section, I introduced two primary constructs of teaching practice, classroom environment and instruction, as well as identifying the types of practices that fall into each category. Research has found that cultural/emotional and management environment practices and cross-subject and subject-specific practices are associated with student outcomes. In the next section, I discuss the literature examining the development of beginning teachers' practices in each of these major areas.

Development of Beginning Teachers' Practices

Beginning teachers often struggle with classroom environment practices like establishing a positive classroom climate and classroom management. Several studies have established that classroom management and discipline are the top concerns of beginning teachers (Britt, 1997; Meister & Melnick, 2003; Veenman 1984, 1987). In a review of 83 empirical studies identifying beginning teacher's challenges, Veenman (1984) found that the top three perceived problems of beginning teachers were (1) classroom discipline, (2) motivating students, and (3) organization of class work. Melnick & Meister (2008) used data from a national sample of beginning teachers to examine how teachers concerns change over time. They found significant differences between beginning and experienced teachers in concerns about classroom management. They suggest that "experience gives teachers the confidence to deal with different behavior issues that occur in the classroom" (p. 52).

New teachers also struggle with choice of academic content, how to teach it, and which resources to use (Grossman, 1992; Kauffman, Johnson, Kardos, Liu, & Peske, 2002). In other words, beginning teachers often lack pedagogical content knowledge. In a study of 50 beginning teachers from all grade levels, Kauffman et al. (2002) found that new teachers struggled to prepare content and materials. In a case study of beginning teachers, Shulman and Colbert (1988) found that beginning teachers struggled with transforming content knowledge into comprehensible lessons while simultaneously dealing with classroom management concerns. In a study of 252 pre-service elementary and secondary teachers, Ball (1990) found that the mathematics knowledge for teaching that the prospective teachers brought with them to the classroom was inadequate for teaching mathematics for understanding. Specifically, Ball used questionnaires and interviews to learn whether pre-service teachers could select and generate appropriate representations for teaching students to divide with fractions. Less than 25% of the pre-service teachers selected the correct response.

Despite recent emphasis on the necessity of subject matter expertise for effective teaching, classroom management has long been the primary concern for most new teachers (Fuller, 1969; Kagan, 1992; Veenman, 1984). A study of beginning teachers in the United Kingdom found that they judged their performance based on classroom management rather than instructional practices or student learning (Oberski, Ford, Higgins, & Fisher, 1999). Surveys have shown that beginning teachers feel more prepared to teach their subject matter than they do to manage a classroom (Boe, Shin, & Cook, 2007). Supports for beginning teachers tend to be aligned with their reported challenges. Mentoring relationships for first-year teachers focus more on classroom management than on content-specific instruction (Hobson, Ashby, Malderez, 2009; Pourdavood, Grob, Clark, Orr, 1999).

Some researchers have suggested that beginning teachers may need to attain a minimum level of classroom management proficiency before they are capable of developing in other areas of their practice (Berliner, 1988; Kagan, 1992). This argument is in line with Fuller's (1969) teacher development model that suggests that beginning teachers move from concerns about self, to concerns about the task of teaching, and, finally, to concerns about their students' learning, and that later concerns cannot emerge until early concerns are resolved (Fuller & Bown, 1975). Yet, others argue that new teachers should simultaneously focus on classroom management and procedures while reflecting and improving on their subject matter knowledge and instruction (Grossman, 1992). The underlying theory behind this argument is that instructional teaching practices are critical for student learning and that well executed instructional practices are key for effective management. Additionally, studies have suggested that progression through both concern stages and type of concerns varies amongst new teachers (Cheny, Krajewski, & Combs, 1992; Ryan, 1986).

Research on beginning teachers' instructional practices has typically employed a case study approach (e.g., Grossman & Thompson, 2008). Few studies have systematically examined variation in teachers' initial quality of teaching practices or examined change in beginning teachers' practices over time. One exception is a recent study that investigated whether and how 24 secondary mathematics, English, and science teachers in the United Kingdom changed in their observed classroom quality during their teacher education year and the first two years of professional practice (Malmberg, Hagger, Burn, Mutton, & Colls, 2010). The teachers showed improvement throughout their first two years of teaching on classroom and time management practices. They also made progress on classroom culture and general instructional practices until about midway through their first year, at which point they declined. The authors suggest that a reduction of supports after the initial entry into teaching may be responsible for the decline.

Factors that might influence the development of beginning teachers' practices include available curricular resources, school environment, and teaching assignments. Inexperienced teachers are especially prone to stick close to the curriculum and materials that they are provided (Grossman & Thompson, 2008), and studies have found that curriculum is connected to quality of instructional practices (Smith, Neergaard, Hochberg, & Desimone, under review). School environment may also impact the development of beginning teachers. Sass et al. (2010) found that new teachers in high-poverty schools improve more slowly than teachers in low-poverty schools. Futhermore, recent research suggests that improvements to teacher effectiveness based on gaining more experience are greater when teaching experience is accumulated in the same grade. Ost (2009) found that teachers who consistently teach the same grade level improve approximately 35% faster than teachers who never repeat grade assignments. Beginning teachers are significantly more likely to switch grades compared to more experienced teachers (Brummet & Gershenson, 2012). It is likely that other changes such as changing schools may also impact the development of beginning teachers' practices.

In summary, research has shown that beginning teachers struggle with both their classroom environment and instructional practices, but that they are most concerned about their management practices. Researchers have debated whether new teachers should first focus on mastering their management practices or work simultaneously to improve both their classroom environment and instructional, yet no large-scale studies have attempted to address this issue using empirical evidence. Just one study was identified that has examined the development of beginning teachers' classroom environment and instructional practices across multiple years. Findings from that study showed improvement in both areas of practice, but more sustained improvement in environment practices (Malmberg et al., 2010). The lack of empirical studies that have examined the growth of beginning teachers' practices also means that little is known about the development of different teaching practices.

Relationship of Teaching Practices to Student Achievement

In the previous sections, I reviewed the literature on two areas of teaching practices, classroom environment and instructional, and the development of beginning teachers' practices in these areas. In order to better understand the relationship of beginning teachers' practices to their effectiveness at increasing their students' achievement, I reviewed prior studies that have investigated the relationship between teaching practices and teachers contributions to student achievement, typically determined by value-added estimates which measure an individual teacher's unique contribution to student achievement.

Overall, research examining the relationship between teaching practices and student achievement tends to find an association between "higher quality" teaching practices and student achievement (Holtzapple 2003; Kane et al., 2010; Milanowski, 2004). However, the meaning of "higher quality" depends on the measure used to capture the teaching practices. The inability to document and examine all the teacher behaviors occurring during a lesson has led researchers to focus their observations on fewer specific behaviors (Schacter & Thum, 2004). After all, as Matsumura et al. (2006) wrote, ". . . it is not feasible to measure all of the skills needed to teach effectively" (p. 5). The studies reviewed use different measures of teaching practices that are based on distinct theories of the practices that represent effective teaching.

I divide the literature examining the relationship between teaching practices and student achievement into four categories based on the practices studied. The first category of studies used combined measures of both classroom environment and instructional practices. The next category focuses on classroom environment practices. The third category focuses specifically on instructional practices and includes both studies of cross-subject instructional practices and studies of subject-specific instructional practices. The final category of research used multiple measures of teaching practice to assess the relationship of teaching practices to teacher effectiveness at increasing student achievement. A table of studies included in this review is available in the Appendix.

Several different types of measures have been used to assess the quality of teaching practices. The most common means for assessing teaching practices are surveys or logs completed by teachers themselves or observational checklists or rubrics completed by outside observers. While surveys and logs can capture a broad range of practices, the practices are self-reported; therefore, observational measures are typically preferred for research studies. Given this preference and since the analysis reported in this paper uses observational measures, this review focuses on studies using observational measures of teaching practices. This review is limited to research published in the last decade when a number of observation protocols rating teaching practices have been developed and used in a variety of settings (Grossman et al., 2008). As very few studies have examined the link between teaching practices and student achievement just for beginning teachers, this review covers published studies for teachers across all experience levels.

The research reviewed here guided the methods used in this paper to investigate the relationship between ratings of classroom environment and instructional teaching practices with beginning teachers' effectiveness at increasing student achievement. Information regarding the magnitude of the relationships between teaching practices and teacher effectiveness found in prior studies assists with later interpretation of the findings from this study. These findings provide a helpful background for considering beginning middle school mathematics teachers' initial ratings on a variety of classroom environment and instructional practices.

Combined Measures of Teaching Practices

Studies reviewed in this section used an overall measure of teaching practice that covers both classroom environment and instructional practices. The measures typically evaluated both cultural and social/emotional practices and management practices and cross-subject instructional practices, as these overall measures were designed to rate teachers across subject areas. The studies reviewed here included teachers in the elementary and middle grades across a variety of school districts in the United States. Overall, the research examining the relationship between overall measures of teaching practices and student achievement typically finds an association between "higher quality" teaching practices and student achievement, but the definition of "higher quality" practices varied depending on the measure used.

Many of the studies relied on data gathered from teacher evaluation systems based on Danielson's (1996) Framework for Teaching (FFT). The FFT is a cross-subject measure that uses a 4-point scale to rate teachers on four domains: Planning and Preparation, Classroom Environment, Instruction (formerly called Teaching for Learning), and Professional Responsibilities. Scores on the four domains were typically added to yield a composite evaluation score, which provided an overall indicator of teacher performance (e.g., Holtzapple, 2003; Milanowski, 2004), but sometimes the four scores were averaged to get an overall measure of teaching quality (e.g., Borman & Kimball, 2004). Another cross-subject measure that has been used to assess teaching practices is the Classroom Assessment Scoring System (CLASS). The secondary school version of the CLASS (CLASS-S) measures teachers' instructional quality across content areas in three broad domains: (a) Emotional Support (ES), (b) Classroom Organization (CO), and (c) Instructional Support (IS) (Pianta, Hamre, & Mintz, 2011). The FFT's Classroom Environment domain and the CLASS-S Emotional Support and Classroom Organization domains assess classroom environment practices. The FFT's Instruction domain and the CLASS-S's Instructional Support domain both assess cross-subject instructional practices.

Holtzapple (2003) linked teachers' ratings from the Cincinnati Public Schools' teacher evaluation system, which was based on standards adapted from the FFT, to student achievement gains on reading, mathematics, science, and social studies tests. Ratings were based on six classroom observations and review of a portfolio including artifacts such as parent contact logs, lesson and unit plans, and examples of student work. Analysis was conducted at the classroom level for about 80 teachers of grades 3 through 8 for the 2000-01 school year and 166 teachers of grades 3, 4, 6, 7, and 8 for the 2001-02 school year. Correlations between the sum of the four domain scores and student achievement ranged from 0.26 (science) to 0.38 (mathematics).

Like Holtzapple (2003), Milanowski (2004) used data from the Cincinnati evaluation system to analyze the relationship between teacher evaluation and value-added scores in the 2000-01 and 2001-02 school years as well. The sample included 212 teachers in grades 3–8. He found small to moderate correlations between average teacher evaluation and value-added scores. The average correlations were 0.27 in science, 0.32 in reading, and 0.43 in mathematics.

Borman and Kimball (2005) also used data from a teacher evaluation system built on Danielson’s (1996) framework. They linked 2002-03 evaluation data from about 400 fourth, fifth, and sixth grade teachers in a Nevada school district to student achievement scores from reading and mathematics district and state assessments. Analyses were conducted at the classroom level using hierarchical linear modeling to predict classroom mean achievement based on teacher’s evaluation scores from four standards from the Instruction and Planning and Preparing domains. While controlling for teacher experience and students’ prior achievement, minority, and free lunch status, they found that a teacher with an evaluation score one standard deviation above the mean was significantly associated with average classroom achievement scores one fifth of a standard deviation above scores of students taught by a teacher at one standard deviation below the mean.

Heneman, Milanowski, Kimball, and Odden (2006) assessed the relationship between student achievement and teachers’ evaluation scores using data from four evaluation systems encompassing all four FFT domains. The four sites were in Cincinnati; Los Angeles; Reno/Sparks, Nevada; and Coventry, Rhode Island. The authors used value-added estimates based on prior achievement and other student characteristics. They found positive correlations between teacher evaluation scores and value-added scores, with variability across location. Correlations ranged from 0.22 to 0.37 for reading and 0.11 to 0.32 for mathematics. The authors also found a fairly high correlation in two of the schools between teachers’ observed practices and their students’ achievement gains. They hypothesized that using multiple, highly trained evaluators led to the higher correlations. They also reported that the sites with higher correlations had a shared understanding of good teaching practices.

Kane et al. (2010) used data from the Cincinnati Public Schools evaluation system based on the FFT. They used data from school years 2000-01 to 2008-09 to test whether ratings of classroom observations identified teaching practices most likely to raise achievement. They focused on the eight standards representing the Classroom Environment and Instruction domains. Kane et al. used a principal components factor analysis to identify three distinct areas of practice: (a) an average of all eight practices, (b) classroom environment score minus instruction score, and (c) teaching through questioning and discussion minus practices related to routinized standards and content-focused teaching. They constructed value-added estimates, then divided teachers into quartiles of their valued-added scores then conducted mean difference tests (t-tests) between the evaluation scores of (1) teachers in the upper quartile of value-added compared to those in the lowest value-added quartile and (2) upper quartile teachers compared to teachers in the second quartile. The researchers found statistically significant mean differences between evaluation scores for the teachers in the higher value-added quartile, but they do not report on the magnitude of these differences. Next, Kane et al. (2010) used regression analysis to examine the extent to which the ratings of the three practice areas were associated student achievement growth. They controlled for student prior achievement and other observable characteristics and used year fixed effects to control for differences in the rubrics over time. A one point increase in the average score across the eight standards was associated with a student achievement gain of about one-sixth of a standard deviation in mathematics and one-fifth of a standard deviation in reading. A one point increase in the average scores represented an increase of about two standard deviations.

Allen, Pianta, Gregory, Mikami, & Lun (2011) conducted a randomized controlled trial of My Teaching Partner, a web-mediated professional development approach focused on improving teacher-student interactions. They examined whether the training led to student achievement gains and whether those gains were mediated by changes to teaching practices, measured by the CLASS-S. My Teaching Partner – Secondary includes workshop-based training, a video library, and a year of personalized coaching in which coaches watch recordings submitted by the teachers and "illustrate either positive teacher interactions or areas for growth in one of the dimensions in the CLASS-S" (p.1035). Their study included 78 secondary teachers in Virginia that were randomly assigned to receive the training. They found gains in student achievement on the state test in the year following the completion of the training. These gains appeared to be mediated by changes in teaching practices that were targeted by the training. They assessed the meditating role of teaching practices using multilevel structural equation modeling. No evidence indicated that the effectiveness of the training depended on the subject taught by the teacher.

Rather than using an existing tool, Schacter and Thum (2004) created their own rubrics to evaluate several teaching practices on five performance levels. The instructional teaching practices included were questions, feedback, presentation, lesson structure and pacing, lesson objectives, thinking, and activities. The classroom environment practices included were classroom environment, grouping students, motivating students, and teacher knowledge of students. These practices were selected based on prior research that randomly assigned teachers to training based on teaching models and found large effect sizes (d = 0.46 – 1.53) in reading, language, mathematics, and social science (Gage & Needles, 1989). Schacter and Thum used the rubrics to rate the practices of 52 teachers from five elementary schools during eight observations in the 2001-02 school year. To investigate the relationship between the ratings of teaching practice and student achievement, they used regression analysis conducted at the classroom level. Findings revealed that having a higher average rating of teaching practices was highly predictive of student achievement gains on standardized achievement tests in reading, language, and mathematics. They also observed correlations between teachers' value-added scores and the average rating of observed teaching performance of 0.55 for mathematics, 0.68 for reading, and to 0.70 for language.

Classroom Environment Practices

Few studies have focused specifically on classroom environment practices. In fact, just two of the studies reviewed here not using multiple measures focused specifically on classroom environment practices. Both of these studies, conducted in grades 3-8 math and reading and grade 9 Algebra, found evidence that classroom environment practices are significantly associated with increased teacher effectiveness.

In addition to using an overall average of the FFT's Classroom Environment and Instruction practices, Kane et al. (2010) also focused on a classroom environment score. This score was based on teachers' practices creating an environment of respect and rapport, establishing a culture for learning, and managing classroom procedures, student behavior, and physical space. Their results also showed that holding average scores constant, higher classroom environment practices were predicted to generate additional student gains in mathematics (one-fourth of an SD) and reading (one-seventh of an SD).

Bell et al. (2012) examined the relationship between teaching practices, measured by the CLASS, and student learning, while framing an approach to build a validity argument for observation protocols. They explained how data from observation scores, value-added models, generalizability studies, and measures of teacher knowledge, student achievement, and teacher and student beliefs can be used to establish validity for observation instruments. They illustrated this approach using data from observations of 82 Algebra teachers and their value-added estimates based on student scores from an algebra end of course (EOC) exam. Bell et al. reported that across the three CLASS-S dimensions, teachers scored highest on Classroom Organization (5.67) and lower on Emotional Support (4.00) and Instructional Support (3.61). The correlation between teachers' value-added scores and Classroom Organization ratings was statistically significant at 0.25, which was the highest of all the correlations. The correlation for Emotional Support was 0.20.

Instructional Practices

A few of the studies that used combined measures of classroom environment and instructional practices also included separate analysis focused just on cross-subject instructional practices. While both Holtzapple (2003) and Kane et al. (2010) reported finding a relationship between reading value-added scores and ratings from the cross-subject measures, they reported that the ratings from cross-subject tools were not significantly associated with gains in mathematics. Moreover, Bell et al. (2012) reported a smaller correlation with math teachers' value-added scores and instructional practice ratings than classroom environment practices. These findings question the validity of using tools assessing cross-subject instructional practice in mathematics. More details on the studies are provided below.

Holtzapple (2003) conducted a separate analysis focused on the Instruction domain because "its content includes the teacher behaviors most likely to be related to student learning" (p. 211). She reported that teachers who received the lowest rating on the Instruction domain had students who, on average, performed lower on the state and district tests than predicted based on prior year test scores, except in mathematics in 2001-02. Students of teachers who received the highest ratings generally performed better than predicted, especially in mathematics. Like Holtzapple, in addition to a combined measure of practice and a separate measure of classroom environment practices, Kane et al. (2010) also examined the additional impact of having higher ratings of teaching through questioning and discussion They found that teachers who scored higher on teaching through questioning and discussion were predicted to produce achievement gains in reading (one-seventh of an SD) but not mathematics. Bell et al. (2012) also focused on cross-subject measures of instruction practices when reporting correlations between Algebra teachers' value-added scores and their ratings on the CLASS-S Instructional Support domain. The correlation of 0.19 was lower than the correlations with Classroom Organization and Emotional Support ratings.

In the past decade, there have also been a few studies focused specifically on establishing a relationship between subject-specific measures of instructional practice and teacher effectiveness. One of the studies reviewed here evaluates the relationship between both reading-specific and math-specific instructional teaching practices and student learning, and the other focuses specifically on math-specific instructional teaching practices. Though the two studies focused on math use different tools to asses teaching practices and different methods to assess the relationship between practice and student achievement, they both found an association between higher ratings and student achievement. The study using both math-specific and reading-specific tools found a positive relationship between the practice ratings and student achievement for both subjects, but the relationship for the math-specific practices appeared stronger.

Matsumura et al. (2006) rated teachers' math and reading practices and assignments using the Instructional Quality Assessment (IQA) math and reading toolkits. Then, they analyzed the ability of the ratings to predict student learning gains. Both the math and reading versions of the IQA assess the quality of observed classroom instruction based on the rigor of lesson activities and the quality of classroom discussion. This pilot study of the IQA was conducted in five urban middle schools. Due to a small sample size, they used linear regression to explore the relationship between the IQA ratings and achievement scores. After controlling for students' prior achievement and background characteristics, the IQA measure of reading comprehension was a significant predictor of students' achievement on the reading comprehension subscale of a standardized achievement test (β = .09, *p* = .05). The math-specific IQA observation ratings significantly predicted student achievement on the total math subscale (β = .16, *p* = .00), and the procedures subscale (β= .32, *p* = .00). Matsumura et al. do not account for clustering within classrooms and schools which may lead results to appear stronger than they actually are.

Another example of a subject-specific tool is Hill and colleagues' framework for evaluating the quality of math instruction (MQI). The MQI rates teachers on classroom work connected to mathematics, richness of mathematics, errors and imprecision, student participation in meaning making and reasoning, and thoroughness in content presentation. Hill, Kapitula, & Umland (2010) differentiated the MQI from the math-specific IQA by explaining that the IQA focuses on the degree to which instruction matches "reform" ideals, whereas the MQI is "more agnostic with regard to teaching method" or curriculum (p. 12). Hill et al. (2010) examined the correlation of 24 middle school mathematics teachers' value-added scores to MQI ratings. They found that teachers' value-added scores correlated with their MQI ratings. The authors report Spearman rank order correlations ranging from 0.36 to 0.45 between teacher's MQI ratings and student outcomes, measured by three different value-added models. The model that controls for student background characteristics showed a correlation of 0.36 between value-added scores and MQI ratings.

Multiple Measures of Teaching Practices

The Measures of Effective Teaching (MET) project, funded by the Bill & Melinda Gates Foundation, is the largest study conducted that compares multiple observation measures of instructional quality with student achievement gains. However, it is not the only study to use multiple measures of teaching practice. Grossman et al. (2010) examined which classroom practices differentiated middle school English/Language Arts (ELA) teachers with high impact on student achievement from those with lower impact using dimensions from the CLASS Emotional Support and Classroom Organization domains, along with a protocol developed specifically for assessing secondary ELA instruction. The Protocol for Language Arts Teaching Observations (PLATO) rates teachers on their instructional scaffolding through teacher modeling, explicit teaching of ELA strategies, and guided practice.

Grossman et al. (2010) focused on grades 6-8 teachers in their third through fifth years of teaching in New York City. The researchers selected 12 pairs of teachers in the same schools in the second (moderate-performing) and fourth quartiles (high-performing) of value-added scores. Teachers were observed on six separate days during the spring of 2007. In general, teachers were stronger on the practices evaluated by the CLASS than those practices assessed by PLATO when comparing on a 1-7 scale. Across all PLATO elements, the high value-added teachers scored higher than the low value-added teachers. These differences were only statistically different for the element of explicit strategy instruction (2.5 compared to 1.9) and marginally statistically different for guided practice (3.2 compared to 2.7) and intellectual challenge practices (4.0 compared to 3.4). High value-added teachers also received more positive scores on each of the CLASS elements, but only student engagement differences were significantly different (4.8 compared to 3.9).

The MET project also used both the CLASS and PLATO. More than 900 observers rated videos of over 1,000 fourth through eighth grade teachers in six districts using five measures of teaching quality. These include two cross-subject measures, FFT (only Classroom Environment and Instruction domains) and CLASS; PLATO, which is ELA specific; and MQI and the UTeach Teacher Observation Protocol (UTOP), which are both math-specific. Mathematics observations were rated using the FFT, CLASS, MQI, and UTOP; English/Language Arts observations were rated using the FFT, CLASS, and PLATO. All the measures except the UTOP have been described in other studies covered in this review. The UTOP was created to evaluate pre-service math and science teachers. It evaluates teachers on four sections: Classroom Environment, Lesson Structure, Mathematics Content, and Implementation.

In the MET Project's *Gathering Feedback for Teaching* research paper, Kane and Staiger (2012) reported strong correlations across all five measures. They report disattenuated correlations "to distinguish teacher-level correlation in the overall scores, as opposed to other sources of variation coming from rater error or particular lessons being observed" (p. 31). CLASS ratings correlated strongly with MQI ratings (r = 0.69) and UTOP ratings (0.68), but were most strongly correlated with FFT ratings (r = 0.88) and PLATO ratings (r = 0.86). Ratings from the two math-specific measures, the MQI and UTOP, also had a strong correlation of 0.85.

Kane and Staiger (2012) also noted areas of strength and weakness for teachers' practices for each of the observation measures used. Across instruments, teachers rated higher on classroom environment practices compared to instructional practices. Specifically, teachers tended to perform well at behavior management, productivity, and creating an environment of respect and rapport (CLASS, FFT, PLATO). Teachers scored lower in areas such as problem solving (CLASS), effective discussion (FFT), intellectual challenge (PLATO), richness (MQI) and investigation (UTOP).

Using data from state and project-administered assessments, the MET project found that all five observation instruments were positively associated with measurement of student achievement gains. They used gains rather than end-of-year scores to better capture whether progress made by students is related to their teachers' practices. When calculating each student's achievement gain on the state and supplemental tests, they controlled for the individual student's characteristics (including prior state test scores) and the mean characteristics of the students in each classroom (to account for peer effects). Researchers found that students with teachers with observation scores in the top quartile on the CLASS, FFT, or UTOP (above the 75th percentile) moved ahead of comparable students by 1.5 months. In contrast, students whose teachers were in the bottom quartile (below the 25th percentile) in classroom observation scores fell behind comparable students by roughly one month of schooling, as measured on the math achievement test. These differences were about half as much for ELA. However, they found a relationship similar to the one between teacher rating scores and student math test gains when examining student gains on the open-ended reading assessments administered by the project. Table 1 shows differences in means gains by top and bottom quartile classrooms on the measures of teaching practice. The differences in gains are reported in student-level standard deviation units. A difference of 0.25 standard deviations is approximately equivalent to one year of schooling.

MET Project researchers also conducted correlations with teachers' ratings on the measures of teaching practice and teachers' value-added scores. Table 1 includes the correlations of the observational measures of teaching practices and teachers' value-added scores. These correlations provide some evidence for a stronger relationship between student achievement gains and ratings from subject-specific measures of teaching practices compared to ratings from cross-subject tools. Ratings from the UTOP were more highly correlated with teacher value-added in mathematics than the ratings from the FFT and CLASS. However, ratings from the MQI were less correlated with student achievement gains in mathematics than ratings from the FFT and CLASS.

Table 1. Relationship between Student Achievement and Measures of Teaching Practice from MET Project (Kane & Staiger, 2012)

	Mathematics		English/Language Arts	
	Difference between Top and Bottom Quartiles	Correlation with Value-added	Difference between Top and Bottom Quartiles	Correlation with Value-added
CLASS	0.10***	0.24	0.01	0.10
FFT	0.07***	0.19	0.02***	0.11
UTOP	0.07**	0.26		
MQI	0.05**	0.16		
PLATO			0.04**	0.24

Review of Research on Teaching Practices and Student Achievement

In summary, the most common method of analyzing the relationship between teaching practices and teacher effectiveness was correlating observation ratings and teacher value-added scores. Generally these correlations ranged from 0.20-0.40, though Schacter and Thum (2004) found correlations ranging from 0.55-0.70. Several studies also examined this relationship by separating teachers into quartiles of their valued added scores and then conducting mean difference tests (t-tests) between evaluation scores of teachers in the different value-added quartiles. Using this method, researchers found that teachers in the higher quartiles of value-added had higher ratings of practice than teachers in the lower quartiles. This finding was consistent across studies using a variety of measures of teaching practices (Grossman et al., 2010; Kane et al., 2010; Kane & Staiger, 2012). A few studies used regression analysis to predict student achievement or teacher effectiveness based on ratings of teaching practices (e.g., Matsumura, 2006). These studies reported magnitude of predicted increase in student achievement based on increases in practice ratings. For example, Kane et al. (2010) reported a one point increase in average score of teaching practices was associated with a student achievement gain of about one-sixth of a standard deviation in math and one-fifth in reading.

Many of the studies used the different techniques described above to examine differences in the relationship between teaching practices and student achievement by subject area. Interestingly, despite using measures that captured similar practices, the findings were not consistent. Some of the studies found support for stronger relationships between practice ratings and measures of student achievement for reading or language arts (Heneman et al., 2006; Schacter & Thum, 2004), whereas others found evidence of stronger relationships for math (Holtzapple, 2003; Kane et al., 2010; Milanowski, 2004). These inconsistent findings could be due to using an aggregate measure of multiple practices, rather than grouping similar practices.

Few studies have investigated differences in the relationship between student achievement and specific areas of teaching practice. Kane et al. (2010) reported a stronger relationship for classroom environment practices and student achievement than for instructional practices and student achievement. Bell et al. (2012) found a stronger correlation for behavior management and productivity practices, as compared to socio-emotional practices and instructional practices. Only the MET Project collected ratings on both measures of cross-subject and subject-specific instructional practices. However, the MET Project researchers used aggregate scores for each of the five measures rather than focusing on areas of practice, therefore, none of the studies actually compared the relationship between cross-subject and subject-specific instructional practices with student achievement.

Gaps in Previous Research

Studies of teaching practices have established the importance of particular classroom environment and instructional practices. Research examining the development of beginning teachers' practices has found that teachers new to the profession struggle with both classroom environment and instructional practices, but few studies have examined the initial quality of teaching practices for beginning teachers for which types they are more likely to improve. The one study that used ratings of teaching practices to observe the development of beginning teachers' practices used only two years of data from one measure of teaching practices, limiting their ability to examine the connection between improvements in areas of practice (Malmberg et al., 2010). Thus, while research has established that beginning teachers improve in their effectiveness during their first few years on the job (Clotfelter et al., 2007; Harris & Sass, 2007; Kane et al., 2006), we are uncertain as to what changes in their practice may be driving their improvement. As Kane & Staiger (2012) wrote in their MET project report, "We do not know which competencies are most susceptible to improvement" (p. 33).

As previously stated, studies investigating the link between teaching practices and teacher effectiveness at increasing student achievement typically focus on a narrow range of teaching practices or an average rating of teaching quality across a variety of practices. Moreover, few studies have examined the relationship between the quality of classroom environment practices and both cross-subject and subject-specific instructional practices with effectiveness at increasing students achievement for the same group of teachers. Accordingly, we still lack understanding of the relationship between areas of practices and of whether certain practices are more effective for increasing student achievement than others. The connection between teaching practices and teacher effectiveness at increasing student achievement is especially unclear for beginning teachers, as no studies of this relationship have focused on beginning teachers.

I address these gaps in the literature by investigating the initial level and subsequent improvement over three years on classroom environment as well as both cross-subject and mathematics-specific instructional practices for a sample of beginning middle school mathematics teachers. I also assess the relationship between teachers' classroom environment and instructional practices and whether beginning teachers adhering to certain practices were more successful in raising student achievement scores. The following section describes the data and analysis methods used to address the research questions.

CHAPTER III

METHODS

To examine the extent to which beginning teachers improve on different aspects of their classroom environment and instructional practices during their first three years of teaching and the relationship of the quality of their practices to their effectiveness at increasing student achievement, I used teacher and student level data from a longitudinal study of beginning middle school mathematics teachers' induction and mentoring experiences, the Assessing Induction and Mentoring project (AIM).

Participants

Teachers were invited to participate in AIM if they met two inclusion criteria: (1) served as the teacher of record for at least one seventh or eighth grade math class; and (2) had no prior experience as a teacher of record. Stipends were offered for each year of participation. About 50% of eligible teachers who were recruited to the study participated in at least one component of data collection. All AIM participants who were observed at least once are included in this study.

Participants included 62 teachers in 11 districts across 4 states from three cohorts who began teaching in either 2007-08, 2008-09, or 2009-10. Table 2 shows the number of teachers in the study from each district. The districts ranged in size and student composition. About one-third of the teachers in the study were from the largest participating district. The largest district enrolled about 98,000 students, and the smallest enrolled about 7,000 students. Percentage of the districts' students receiving free or reduced price lunch (FRPL) ranged from 9% to 66%.

Table 2. Characteristics of Participating School Districts[1]

State	District	Number of Teachers in Study	Number of District Secondary Teachers	Urbanicity	Schools	Students	Black or Hispanic	FRPL
1	A	12	520	Urban	70	37,000	33%	47%
	B	21	1,480	Urban	180	98,000	41%	59%
	C	3	120	Rural	10	8,000	7%	51%
2	D	2	300	Suburban	20	11,000	14%	14%
3	E	1	190	Suburban	10	7,000	69%	63%
	F	1	430	Suburban	20	12,000	12%	9%
4	G	7	1,420	Urban	140	75,000	63%	66%
	H	4	780	Rural	40	38,000	25%	37%
	I	7	820	Suburban	50	48,000	42%	31%
	J	3	560	Suburban	50	27,000	14%	35%
	K	1	580	Rural	40	31,000	8%	10%

About one-third of the teachers were male, most were white, and all had bachelor’s degrees. Teachers’ academic backgrounds were categorized as math, math education, education, or other based on the teachers’ majors undergraduate or graduate majors. Teachers were categorized as having a math degree if they listed their major or minor area of study for either a bachelor’s or master’s degree as mathematics, and they did not have a math education or education degree. Teachers were classified as having a math education degree if they had a bachelor’s degree with a major or minor in math education, a master’s degree in math education, or a combination of math and education degrees. Teachers were categorized as having an education degree if they had a bachelor’s degree with a major or minor in education or a master’s degree in education without a math or math education degree. The degree categories are mutually exclusive. Teacher background characteristics are shown in Table 3.

I explored the extent to which teachers who participated in the AIM study had similar background characteristics to a nationally representative sample of beginning mathematics

[1] Common Core of Data (2009-2010). Number of schools rounded to the nearest ten and number of students rounded to the nearest thousand to protect the identity of the school districts.

teachers in public middle schools in the 2007-08 Schools and Staffing Survey (SASS). The SASS sample included fulltime first-year middle school math teachers and, for a larger sample, those with five years of teaching experience or less. Compared to SASS first year middle school mathematics teachers (n=44), AIM teachers were similar in age, gender, race, and percentage with alternative certification, but AIM participants were more likely to have math education degrees and less likely to have student teaching experience. None of these differences were statistically significant.

Table 3. Teacher Background Characteristics from AIM and 2007-08 SASS

	AIM Teachers (n=62)	SASS First-Year Teachers (n=44)	SASS Beginning Teachers (n=275)
Age	27.9	27.4	32.2
Male	31%	25%	30%
White	89%	86%	87%
Education Degree	42%	48%	48%
Math Education Degree	24%	14%	15%
Math Degree	8%	5%	7%
Other Degree	26%	34%	30%
Alternative Certification	27%	32%	28%
Student Taught	62%	73%	81%

Measures

Researchers videotaped participating teachers' instruction on two consecutive days during the same class period at four time points: winter of the first year of teaching and the spring of the first, second, and third years. One district did not allow videotaping therefore live rating and audio recording was used. Videographers logged the sequence of activities that occurred in the classroom and collected or recorded (by hand or with video) student assignments.

Table 4 shows the number of recordings that were conducted at each time point. Attrition due to teachers either leaving the profession (n=8), moving to different school districts (n=10), changing subject areas or transferring out of middle grade levels (n=2), or leaving the study (n=7), resulted in a total of 35 teachers having a complete three years of data. All observations were rated using both the math-specific Instructional Quality Assessment (IQA) and the cross-subject Classroom Assessment Scoring System (CLASS). Detailed descriptions of the two tools, the processes used to rate the videos, and the methods used to assess inter-rater reliability are discussed below.

Table 4. Number of Teachers Recorded at Each Time Period

	Cohort 1	Cohort 2	Cohort 3	All Cohorts
Year 1 Fall	23	25	13	61
Year 1 Spring	21	24	14[2]	59
Year 2	14	19[3]	11	44
Year 3	13	15	7	35

Classroom Assessment Scoring System (CLASS). The CLASS observation tool has been used by teacher preparation programs and for teacher performance assessment, professional development, program monitoring, and research and evaluation. The following sections describe the secondary version of the CLASS and the rating process and methods used to assess reliability.

CLASS-S description. The secondary version of the CLASS measures teachers' instructional quality across content areas in three broad domains: (a) Emotional Support (ES), (b) Classroom Organization (CO), and (c) Instructional Support (IS) (Pianta, Hamre, & Mintz, 2011). Both the Emotional Support and Classroom Organization domains assess classroom environment practices. The Emotional Support domain captures cultural and emotional practices, while the

[2] Teachers entered the study late due to switching subject areas and grade levels in January of year one.
[3] One teacher observation is missing because the teacher is out for maternity leave, but this teacher is observed in year 3.

Classroom Organization domain focuses on behavioral and time management practices. Instructional Support captures instructional practices applicable across subject area.

Each domain is organized into multiple dimensions, and each dimension consists of several indicators (see Table 5). The *Emotional Support* domain includes positive climate, negative climate, teacher sensitivity, and regard for adolescent perspective. Positive climate assesses the emotional connections and relationships between teachers and students, and the warmth, respect, and enjoyment communicated by verbal and non-verbal interactions. Negative climate evaluates the overall level of negativity among teachers and students in the class. Teacher sensitivity considers the responsiveness to the academic and social/emotional needs and developmental levels of individual students as well as the entire class. Regard for adolescent perspective measures the extent to which the teacher is able to meet and capitalize on the social and developmental needs and goals of adolescents by providing opportunities for student autonomy and leadership; it also considers the extent to which student ideas and opinions are valued and content is made useful and relevant to adolescents.

The *Classroom Organization* domain includes behavior management, productivity, and instructional learning formats. Behavior management evaluates the teacher’s use of effective methods to encourage desirable behavior and prevent and redirect misbehavior. Productivity considers how well the teacher manages time and routines so that instructional time is maximized. The dimension of instructional learning formats assesses the ways the teacher maximizes student engagement in learning through clear presentation of material, active facilitation, and the provision of interesting and engaging lessons and materials.

The *Instructional Support* domain consists of content understanding, analysis and problem solving, quality of feedback, and instructional dialogue. Content understanding measures both the depth of lesson content and the approaches used to help students comprehend the framework, key ideas, and procedures. Analysis and problem solving assesses the degree to which the teacher facilitates students' use of higher level thinking skills, such as analysis, problem solving, reasoning, and creating through the application of knowledge and skills. Quality of feedback evaluates the way the teacher's feedback expands learning, extends understanding, and encourages student participation. In secondary classrooms, this dimension acknowledges that peers may also provide feedback. Instructional dialogue considers the purposeful use of dialogue (questioning and discussion) by teachers to facilitate students' understanding of content.

Raters scored each dimension as low (1, 2), mid (3, 4, 5), and high (6, 7). General scoring guidelines are presented in Table 6. An example of the quality of feedback dimension rubric is presented in Table 7. It shows how a dimension is broken down into indicators and how each indicator includes a few behavior markers and describes what would be occurring in a classroom for each of low, mid, and high rating levels.

Raters rate multiple short segments of instruction during a class period instead of rating a single class period of instruction in its entirety. The CLASS-S manual suggests that two segments of rating can be completed during a 45-minute class period, that three can be completed during a 90-minute period, and that at least four segments should be obtained to get an accurate picture of a teacher's instructional quality at a particular point in time.

Table 5. Overview of the CLASS-S Domains, Dimensions, and Indicators

Domains	Dimensions	Indicators
Emotional Support	Positive Climate	• Relationships • Positive affect • Positive communications • Respect
	Negative Climate	• Negative affect • Punitive control • Disrespect
	Teacher Sensitivity	• Awareness • Responsiveness to academic and social/emotional needs and cues • Effectiveness in addressing problems • Student comfort
	Regard for Adolescent Perspective	• Flexibility and adolescent focus • Connections to current life • Support for student autonomy and leadership • Meaningful peer interactions
Classroom Organization	Behavior Management	• Clear expectations • Proactive • Effective redirection of misbehavior • Student behavior
	Productivity	• Maximizing learning time • Routines • Transitions • Preparation
	Instructional Learning Formats	• Learning targets/organization • Variety of modalities, strategies, and materials • Active facilitation • Effective engagement
Instructional Support	Content Understanding	• Depth of understanding • Communication of concepts and procedures • Background knowledge and misconceptions • Transmission of content knowledge and procedures • Opportunity for practice of procedures and skills
	Analysis and Problem Solving	• Inquiry and analysis • Opportunities for novel application • Metacognition
	Quality of Feedback	• Feedback loops • Scaffolding • Building on student responses • Encouragement and affirmation
	Instructional Dialogue	• Cumulative content-driven exchanges • Distributed talk • Facilitation strategies
Student Engagement		• Active engagement

Table 6. CLASS-S General Scoring Guidelines (Pianta et al., 2010)

Low		Mid			High	
1	2	3	4	5	6	7
The low range description fits the classroom/ teacher very well. All, or almost all, relevant indicators in the low range are present.	The low range description mostly fits the classroom/ teacher but there are one or two indicators that are in the mid range.	The mid range description mostly fits the classroom/ teacher, but there are one or two indicators in the low range.	The mid range description mostly fits the classroom/ teacher very well. All, or almost all, relevant indicators in the mid range are present.	The mid range description mostly fits the classroom/ teacher, but there are one or two indicators in the high range.	The high range description mostly fits the classroom/ teacher, but there are one or two indicators in the mid range.	The high rang description fits the classroom/ teacher very well. All, or almost all, relevant indicators in the high range are present.

Table 7. CLASS-S Quality of Feedback Rubric (Pianta et al., 2010)

	Low (1,2)	**Mid (3,4,5)**	**High (6,7)**
Feedback loops • Back-and-forth exchanges • Persistence • Follow-up questions	Feedback in this classroom is non-existent or perfunctory.	There are occasional feedback loops between the teacher and students or among students, but at other times feedback is more perfunctory.	There are frequent feedback loops between the teacher and students or among students, which lead students to obtain a deeper understanding of material and concepts.
Scaffolding • Assistance • Hints • Prompting completion and thought processes	Students are not provided with assistant, hints, or prompting from the teacher or peers when participating in class work but are left to complete work without such assistance.	The teacher and/or peers sometimes scaffold student learning but these interactions are brief or not of sufficient depth to allow students to fully perform at a higher level.	The teacher and/or peers often scaffold student learning, allowing them to perform at a higher level than they would be able to perform independently.
Building on student responses • Expansion • Clarification • Specific feedback	The teacher and/or peers move on quickly after a student has provided an answer or presented work without building on student responses in a way that clarifies or extends learning.	The teacher and/or peers sometimes build on student responses to expand students' learning and understanding, but these exchanges are brief and/or limited in depth.	The teacher and/or peers often build on student responses in a way that expands students' understanding.
Encouragement and affirmation • Recognition and affirmation of effort • Encouragement of persistence	Students rarely receive encouragement or affirmation of their work or participation.	The teacher and other students occasionally offer encouragement of students' efforts that increases involvement and persistence.	The teacher and other students often offer encouragement of students' efforts that increases involvement and persistence.

CLASS-S rating and reliability. Three raters used the CLASS-S to rate the classroom lessons. All three CLASS raters were former teachers. Raters participated in either in-person or online training and were certified by Teachstone, a company that oversees the training of the CLASS observation tools. Certification requires completion of a reliability exercise which involves watching several classroom observation segments and rating within one-point of expert ratings on 80% of the dimensions overall. Before the rating team individually rated videos, all three watched six videos (two videos for each of three teachers) and came together to discuss any discrepancies in their ratings.

As the CLASS-S rating system is based on rating short segments of classroom interactions, guidelines were established to divide the recordings into segments. If the video was 65 minutes or less, raters divided the video into two even segments. If the video was longer than 65 minutes, raters divided the video into three equivalent segments. Per the CLASS-S guidelines, raters watched the recordings of classroom instruction for each segment while taking notes on the CLASS-S indicators. Then they took about 10-15 minutes to review their manual and assign scores to each domain. Scoring was completed immediately after each observation was watched. CLASS-S ratings were averaged across segments and then across the two consecutive days for each of the four time points to provide a picture of teachers' instructional quality at a particular point in time.

Acceptable reliability on the CLASS-S is defined as being within one point for 80% of the ratings (Pianta et al., 2011). With 12 possible dimension scores for each segment rated, raters should be within one point of each other on 10 of the 12 dimensions. To assess ongoing inter-rater reliability, 20% of the videos were randomly selected for double rating. While having two raters rate every recording can help reduce possible rater bias and ensure accuracy, this process is time consuming and costly. Since CLASS-S rating took place over a few months and all three raters had passed a reliability exercise to be certified, potential rater bias was less of a concern. Additionally, rather than averaging the ratings of the videos that were double rated, consensus rating was used. During most weeks of the rating process, two of the three raters (on a rotating schedule) were assigned to rate the same teachers' videos from a particular time point. The two raters then met to discuss rating discrepancies and arrive at a consensus. The consensus codes were entered as the final codes. If at any point the two raters did not meet the 80% one-point reliability across all segments rated, the third rater was asked to rate the videos and attend a consensus rating meeting. Consensus rating was not only a means of assessing inter-rater reliability, but the frequent consensus sessions also served as an ongoing means of ensuring a shared understanding of the measures (Stein, Grover, Hennigsen, 1996). On the double coded segments, overall exact-point reliability was 44% and, one-point reliability was 87%.

Mathematics Instructional Quality Assessment (IQA). While the CLASS-S captures a broad range of instructional practices, the Instructional Quality Assessment is specifically geared toward mathematics instruction (Junker et al., 2006; Matsumura et al., 2006; Matsumura, Garnier, Slater, & Boston, 2008). The Learning Research and Development Center at the University of Pittsburg designed the IQA based on specific guidelines for instructional practice articulated in the National Research Council's publication, *How People Learn* (Bransford, Brown, & Cocking, 2000). IQA ratings are based on the degree to which the teacher selects and implements cognitively demanding problem-solving tasks, and organizes discussions emphasizing reasoning and connections among mathematical ideas. When scheduling observations, teachers were told that the researchers were interested in seeing a lesson where students were presented with a problem, had time to explore and work on the problem, and had the opportunity to discuss the problem.

The focus on these math-specific practices is justified by current emphasis on the use of genuine, challenging tasks and classroom discourse that focuses on key mathematical ideas by the National Council of Teachers of Mathematics (2000). Tasks with high cognitive demand require students to make connections to the underlying mathematical ideas, use procedures to solve tasks that are open with regard to which procedures to use, and, ideally, engage students in disciplinary activities of explanation, justification, and generalization. Research has found that assigning tasks of high cognitive demand and maintaining that level of demand throughout the lesson was related to greater student gains on an assessment requiring high levels of mathematical thinking and reasoning (Stein & Lane, 1996). These instructional practices differ from typical math instruction requiring low cognitive demand that require students to memorize or reproduce facts, or perform relatively routine procedures without making connections to the underlying mathematical ideas (Valli, Croniger, & Buese, 2012). The following sections describe the IQA, the rating process, and methods used to ensure reliability.

Math-specific IQA description. The IQA assesses the quality of observed classroom instruction on academic rigor (AR) and accountable talk (AT). IQA ratings range from 0-4, with a 3 or 4 indicating higher levels of instructional quality. The three rubrics assessing AR are (1) task potential, (2) task implementation, and (3) class discussion following the task. The task potential rubric asks: *Did the task have potential to engage students in rigorous thinking about challenging content?* The task implementation rubric focuses on what is occurring during the problem solving activity and centers on the question: *At what level did the teacher guide students to engage with the task in implementation?* For task potential and implementation, a score of 0 indicates absence of mathematical activity; 1 indicates instruction emphasizing facts and memorization; 2 indicates instruction emphasizing unambiguous application of procedures and single representations of concepts; and 3 and 4 designate instruction characterized by open-ended tasks, multiple representation of mathematical concepts, and connections among mathematical ideas, with a 4 awarded if there is explicit evidence of students' reasoning and understanding. The discussion rubric addresses the question: *To what extent did students show their work and explain their thinking about the important mathematical content?* A score of 0 indicates no discussion of the task; 1 specifies discussion where students provide brief or one-word answers; 2 designates discussion in which students show or describe their work for solving the task, but do not talk about their strategies or the mathematical ideas behind the task; and 3 and 4 designate discussion characterized by student explanations for their strategies used to solve the task and connections to the underlying mathematical ideas involved in the task, with a 4 assigned to discussions in which students explain why certain strategies were used.

Accountable Talk (AT) is assessed using five rubrics: (a) participation, (b) teachers' linking, (c) students' linking, (d) teacher asking, and (e) student providing. *Participation* is judged on the percentage of students who participate in the teacher-facilitated discussion following the activity. *Teacher's linking* is rated on the extent to which the teacher supports students in connecting their ideas and positions to each other. *Students' linking* assesses the degree to which students' contributions link to and build on each other. *Teacher asking* looks at whether the teacher presses students to support their contributions with evidence and reasoning. *Student providing* rates the degree to which students supported their contributions with evidence and reasoning. Table 8 provides an overview of the IQA rubrics and their guiding questions, and

Table 9 details the scoring guidelines for task potential and implementation and student discussion.

Table 8. Overview of the Math-specific IQA Rubrics and Guiding Questions

	Rubrics	Guiding Question
Academic Rigor	Task Potential	Did the task have potential to engage students in rigorous thinking about challenging content?
	Task Implementation	At what level did the teacher guide students to engage with the task in implementation?
	Student Discussion Following Task	To what extent did students show their work and explain their thinking about the important mathematical content?
Accountable Talk	Participation	Was there widespread participation in teacher-facilitated discussion?
	Teacher's Linking	Does the teacher support students in connecting ideas and positions to build coherence in the discussion?
	Students' Linking	Do student's contributions link to and build on each other?
	Asking (Teachers)	Were students pressed to support their contributions with evidence and/or reasoning?
	Providing (Students)	Did students support their contributions with evidence and/or reasoning?

Table 9. Math-specific IQA Scoring Guidelines

	Level	Task Potential and Implementation	Student Discussion
High	4	Explicit evidence of students' reasoning and understanding	Complete and thorough explanations of why strategies, ideas, or procedures are valid; Connections made to the underlying mathematical ideas
	3	Open-ended tasks, multiple representation of mathematical concepts, and connections among mathematical ideas	Explanations of why strategies, ideas, or procedures are valid and/or begin to make connections BUT the explanations and connections are not complete and thorough
Low	2	Procedures and single representations of concepts	Written work for solving the task is shown or described, but no discussion of why strategies, procedures, or mathematical ideas are valid.
	1	Facts and memorization	Brief or one-word answers
	0	Absence of math activity	No discussion

The analysis in this study focused on the three AR rubrics and an overall measure of instructional quality that was created by (1) averaging the task potential and implementation scores, (2) averaging the discussion and five additional accountable talk scores (i.e., participation, teacher linking, student linking, asking, and providing), and (3) averaging the two averages. This overall instructional quality measure was supported by an exploratory factor analysis that indicated that the eight IQA scores separate into two main factors: one that includes the task potential and implementation and another that includes all of the discussion ratings. As all the IQA rubrics were focused on math instructional practices, the overall measure provides a rating of math-specific instructional practices.

Math-specific IQA rating and reliability. Recordings of each lesson were viewed and rated on the IQA by two independent raters from a team, who participated in at least two full days of training conducted by IQA developers. One set of IQA ratings for each of the four observation points was generated by averaging across the two raters for each day and then across the consecutive days as a means of improving reliability. A third rater was brought in if any of the eight rubric scores differed across the two initial raters by two or more points and if the difference crossed the 3-point threshold. For example, a third rater was used if the ratings were 1 and 3, 2 and 4, 0 and 3, but was not used if the ratings were 0 and 2. Codes were then averaged across the two closest raters. A third rater was needed for about 25% of the observations.

Unlike the CLASS-S ratings which took place after the multi-year AIM study was complete, IQA ratings took place in each summer following the school year when the data was collected. Sixteen raters assisted with the IQA rating with the majority completed by 5 raters. Of the 16 raters, 14 had teaching experience with most having taught either middle or high school mathematics. Raters did not have personal relationships with the teachers being observed. Before each round, all raters participated in inter-rater reliability exercises, in which they viewed the same subset of classroom videos, rated the videos, and discussed their ratings to reach mutual understanding and agreement. Raters did not begin individually rating lessons until 80% inter-rater agreement was reached for the exercises. Raters had to be within one point of each other on 7 of the 8 rubrics.

Multiple methods were used to assess the reliability of IQA ratings. Exact agreement between paired raters—calculated as the total number of agreements divided by the total number of agreements and disagreements—was 60%. This is lower than the 81.8% overall exact point inter-rater reliability found in a pilot study of the IQA of 13 middle school mathematics teachers (Matsumura et al., 2006), but higher than the about 50% exact inter-rater agreement on the AR rubrics found in a pilot conducted with 14 elementary school teachers (Boston & Wolf, 2006). One-point agreement, where raters are considered in agreement if individual scores were within one point on each IQA rating scale, was 95% in the elementary pilot but was not reported for the middle school pilot. In our study, one-point agreement was 88%. In addition, a generalizability study (G-study) was conducted at four time points during the rating process to verify that the design for rating lessons—with two raters rating two lessons each at each time point—provided a stable estimate of instructional quality. In these G-studies, each rater involved in the project at the time of the G-study, independently rated two lessons from each teacher in a random sample, and the ratings were analyzed using GENOVA software (Crick & Brennan, 1983). At the teacher level, generalizability coefficients with two raters and two observations for each teacher ranged from 0.74 to 0.98, with an average of 0.81, indicating sufficient reliability. Each time a G-study was conducted, the team of raters met to discuss any discrepant ratings and to come to consensus on the rating scheme.

Comparison of Math-specific IQA and CLASS-S. While the IQA and the CLASS-S are both concerned with the quality of a teacher's instructional practices, they emphasize different criteria. The IQA focuses on elements the developers felt were critical to developing students' conceptual understanding of mathematics. These elements are the rigor of the task provided, student engagement with the task during work time, and discussion following work time. In contrast, two-thirds of CLASS-S dimensions are dedicated to elements of classroom climate and organization, while one-third describes general instructional strategies.

The CLASS-S's Instructional Support domain has some similarities with the concepts measured by the IQA. For example, scoring "high" on the analysis and problem solving dimension is characterized by the following behaviors: the teacher provides opportunities for students to independently solve or reason through novel and open-ended tasks; students consistently engaged in extended opportunities to use higher-order thinking; and students have multiple opportunities to think about their own thinking through explanations, self-evaluations, reflection, and planning (Pianta et al., 2011, p. 76). These indicators are similar to the criteria for scoring a "4" on the IQA's task potential and implementation rubrics. Further, the CLASS-S content understanding dimension, considers "high" instruction as occurring when the "focus of the class is on encouraging deep understanding of content through the provision of meaningful, interactive discussion and explanation" (Pianta et al., 2011, p. 64). This is akin to the condition of students engaging in a discussion of important math ideas for scoring a "4" on the IQA discussion rubric.

Additionally, similar instructional practices are measured by the CLASS-S quality of feedback and instructional dialogue dimensions and the IQA Accountable Talk rubrics. For example, the CLASS dimensions consider instruction to be "high" if the instruction contains "frequent feedback loops . . . which lead students to obtain a deeper understanding of material and concepts" and "content-driven dialogues" (Pianta et al., 2011), and the Accountable Talk rubrics also rate instruction highly if teachers press students to support their contributions with evidence and support students in connecting ideas to build coherence in the discussion. Another example is the IQA participation rubric coherence with both the quality of feedback and instructional dialogue dimensions that look at the ways teachers encourage student participation and whether that participation is distributed among the students. Table 10 shows which IQA rubrics and CLASS-S dimensions measure similar concepts, though it is important to note that there are several aspects of instructional practice assessed by the CLASS-S that are not measured by the math-specific IQA and vice versa. All of the CLASS-S dimensions are from the Instructional Support domain. There is no conceptual overlap between the IQA and the CLASS-S Emotional Support and Classroom Organization domains.

Table 10. Alignment between Math-specific IQA Rubrics and CLASS-S Instructional Support Dimensions

IQA Rubrics	CLASS-S Instructional Support Dimensions and Indicators
Task Potential to engage students in rigorous thinking about challenging content	Analysis and Problem Solving • Opportunities for novel application – teacher provides open-ended tasks and presents cognitive challenges
Implementation of the task	Analysis and Problem Solving • Inquiry and analysis – opportunities for students to use higher-order thinking through inquiry and analysis • Opportunities for novel application – opportunities for students to independently solve tasks • Metacognition – opportunities for students to think about their own thinking
Student **Discussion** of their work and thinking about the content	Content Understanding • Depth of understanding – discussion that emphasizes meaningful relationships among facts, skills, and concepts Quality of Feedback • Feedback loops – feedback loops that lead to deeper content understanding Instructional Dialogue • Cumulative content-driven exchanges – content-driven dialogues that further content knowledge or skills
Participation in teacher-facilitated discussion	Quality of Feedback • Encouragement and affirmation – encouragement that increases involvement and persistence Instructional Dialogue • Distributed talk – balance of teacher and student talk that includes the majority of students and student-initiated dialogues
Teacher Linking of connections between math ideas to build coherence	Content Understanding • Background knowledge and misconceptions – new concepts are linked to students' prior knowledge to advance understanding Quality of Feedback • Building on student responses – teacher builds on responses in a way that expands students' understanding Instructional Dialogue • Cumulative content-driven exchanges – connection to content and exchanges that build on one another
Student Linking to build on ideas	Quality of Feedback • Building on student responses – peers build on responses in a way that expands students' understanding Instructional Dialogue • Cumulative content-driven exchanges - exchanges that build on one another
Teacher Press of students to support their contributions with evidence and reasoning	Quality of Feedback • Scaffolding – assistance, hints, and prompting provided to scaffold learning Instructional Dialogue • Facilitation strategies – open-ended questions and statements
Student Providing support for their contributions	Instructional Dialogue • Facilitation strategies – open-ended questions and statements

Teacher Value-added Scores. Student mathematics achievement data was used to calculate individual teacher value-added scores. These value-added scores are a measure of the teachers' contributions to student achievement in mathematics (Raudenbush & Bryk, 2002; Sanders & Rivers, 1996).

AIM's district partners provided math achievement scores for students whose teachers participated in the study. We requested mathematics scale scores on the state assessments for all students in the math classes taught by the participating teacher for the year the students were in the participating teacher's class and for the two years prior. For example, if a student was in the teacher's class in 2007-08, their 2007-08, 2006-07, and 2005-06 mathematics state test scores were requested. Means and standard deviations of the math scale scores for each grade at the state level for each year were used to standardize the test scores so they were comparable across districts. Test scores were normalized to the state-level because of the compressed variance at the district level. Achievement data was collected from 10 of the 11 participating districts. The district that did not supply data had two teachers that participated in the study. Additionally, achievement data was only collected for cohort 3 teachers for their first two years of teaching.[4]

In each state, multiple choice assessments were administered in the spring and were aligned to state content standards and state performance indicators. As this study includes teachers from multiple states, student scores were from multiple tests with differing scales. Therefore, scale scores for students in each state were placed on a common scale through linear transformation (see Rock et al., 1985), using the state mean to standardize the scores. Linear equating methods (e.g., creating z-scores) create comparable indexes of achievement across

[4] Data collection for the AIM study originally planned to include only two years for Cohort 3 teachers. A third year of classroom observations was added toward the end of the study, but the additional year of student achievement data was not included in this extension.

states (Hedges & Nowel, 1999) and have been shown to perform acceptably with comparable tests (Kolen & Brennan, 1995; Petersen, Cook, & Stocking, 1983).

Districts also provided us with math course title and class period for each student and demographic information for the students including gender, race/ethnicity, free or reduced price lunch status, Limited English Proficiency (LEP) status or English as a Second Language (ESL) status and special education status. These student-level covariates were included as controls in the value-added models because not including them may lead to biased estimates (Amrein-Beardsley, 2008; Kupermintz, 2003).

To calculate value-added scores for teachers, this study uses a model adapted from Henry et al.'s (2011) study using value-added modeling to estimate teacher effectiveness of beginning teachers. Three-level hierarchical linear models where students were nested in teachers in schools were used to estimate teacher effectiveness ratings. The equation below is a reduced form equation for the estimation of student achievement. Teacher-level residuals are recovered and used as estimates of teachers' value-added. The residuals provide an estimate of the extent to which the students' predicted achievement is above or below what would have been expected given their prior achievement and demographic and school characteristics. This variation from predicted achievement is considered to be the teachers' contribution to student achievement.

$$A_{ikt} = \beta_0 + \beta_1 A_{ik,t-1} + \beta_2 A_{ik,t-2} + \gamma_3 SC_{it} + \alpha_{it} + e_{it}$$

The equation predicts the math achievement in school year *t* of student i taught by teacher k in school s, as a function of two years of prior math achievement and student characteristics that may influence achievement.[5] SC_{it} is a vector of student-level covariates including free and reduced price lunch status, race, ESL/LEP status, and special education status.

[5] Students were excluded if they had missing values for the previous year's math achievement score.

A district fixed effect, α_{it}, was also included. The residual or error term, which represents the unexplained variance at the student and classroom level, is used as the value-added measure (VAM). Each teacher's scores are considered to be independent by year so the models were run separately for each year. When running the value-added models, student achievement data is used from all the teacher's math classes in a given year, not just the one observed class. This is a common approach used in studies examining the relationship between ratings of teacher practice and value-added scores (e.g., Hill et al., 2011).

Less than one percent of students were missing outcome scores (i.e., scores for the year they were in the classroom of a teacher in the study). These students are not included in the teacher value-added estimates. About 11% of students were missing a prior year score, and about 24% were missing scores from two years prior. Multiple imputation at the year, state, and grade level was used to impute the missing prior achievement scores.[6] The missing scores were imputed before the scores were standardized. Additionally, one state did not administer a test in sixth grade in 2006 meaning that eighth grade students from this state in 2008 were all missing a score from two years prior to being included in this study. Therefore, only one year of prior data was included for these students. Across years, models with one year and two years of prior generated value-added scores that were highly correlated (r=0.99) so this is not a large concern.

Characteristics of classroom peers are sometimes included as controls in models estimating value-added scores (e.g., Henry, 2011) because the prior achievement of one's peers has been found to influence student achievement (Hanushek, Kain, Markman, & Rivkin, 2003). However, one district did not provide any class period information and another only provided class period data for one year of the study so class characteristics were not able to be calculated for these students. As class prior achievement could not be included for all students, the primary

[6] Multivariate normal regression was used to impute the continuous test scores (Little & Rubin, 2002).

value-added model does not include this variable. As a sensitivity test, I calculated the value-added models including average prior achievement of classroom peers (excluding oneself). For students with class period data, value-added scores resulting from the models with and without peer prior achievement were highly correlated (r=.89). This high correlation indicates that similar value-added estimates resulted from the models with and without class prior achievement. Therefore, I used the value-added scores without peer prior achievement so students without class period information could be included in the primary analysis.

Analysis Methods

The following section describes the analysis methods that I used to (a) examine the initial levels of beginning teachers' classroom environment and instructional practices and the extent to which they improved these practices during their first three years of teaching, (b) investigate the relationship between measures of teachers' environment and instructional practices, and (c) examine the relationship of classroom environment and instructional teaching practices with beginning teachers' effectiveness at increasing student learning.

Development of beginning teacher practices. First, to appraise the initial levels of teaching practices, I presented the average scores of the CLASS-S domain scores and the rubric scores for the math-specific IQA for the fall of the teachers' first year of teaching. I cannot simply compare the average scores because the CLASS-S and the IQA measure different aspects of teaching practice on different scales. A score of 3 on the CLASS-S is not the same as a score of 3 on the IQA. Instead, I used the rating categories of low, mid, and high to conceptualize the mastery of teaching practices demonstrated by the beginning math teachers in this study. Both the CLASS-S and the IQA associated scores with a level of practice. Having a "high" level of practices on the CLASS-S is a score of 6 or 7, and for the IQA it is a score of 3 or 4. Each measure also distinguishes a “middle” category of scores (3, 4, and 5 on the CLASS-S and 2 on the IQA), and a low range of scores (1 and 2 on the CLASS and 0 and 1 on IQA).

I also compared the average initial practice ratings from the teachers in this study to average ratings found in other studies using the CLASS-S and the IQA to rate teaching practices. This provides an understanding of how the initial levels of the beginning middle school math teachers’ classroom environment and instructional practices contrast with the average ratings of these practices for teachers across subjects and experience levels.

To examine to what extent beginning teachers improved their classroom environment and instructional teaching practices, I used growth curve analysis. Growth curve analysis provides an estimate of the amount of predicted growth, as well as an indicator of whether the change over time is statistically significant. Growth curve analysis has rarely been used to examine ratings of teacher’s instructional quality, though it makes sense to expect improvement with experience, especially in teachers’ beginning years. A few studies have used growth curve analysis to examine change over time in university professors’ instructional ratings (Lang & Kersting, 2007; Marsh, 2007). Advantages of individual growth curve modeling are that it captures the time-ordered nature of the observations and assessment times do not have to be identical, therefore allowing respondents with missing data to remain in the analysis (Raudenbush & Bryk, 2002; Singer & Willett, 2003). As the four observation points were not evenly spaced, the month of IQA observation (0=August of first year, 12=August of second year, etc.) was used as the time variable.

I modeled the teachers’ IQA and CLASS-S score trajectories across the four time periods using a multilevel approach to growth curve modeling (Raudenbush & Bryk, 2002; Rogosa, Floden & Willett, 1984). For each of the IQA and CLASS-S outcomes, I ran a two-level hierarchical model allowing for a random slope on intercept and month. In Level 1, Y_{ti} is the IQA or CLASS-S score at month *t* for teacher *i*. The coefficient π_{0i} represents the predicted initial IQA or CLASS-S score for teacher *i* at the start of the first year of teaching (estimated initial score), and π_{1i} is the monthly growth rate for teacher *i*. The e_{ti} is the within-teacher error term, which I assume is normally distributed with a mean of zero and a constant variance. The level 1 model is below.

$$Y_{ti} = \pi_{0i} + \pi_{1i}(Month\ of\ Observation)_{1i} + e_{ti}$$

The level 2 model looks at differences between teachers in their initial status and rate of change. In this model, π_{0i} is the teacher's initial score at the start of the first year of teaching and π_{1i} is the teacher's predicted rate of growth. β_{00} represents the mean initial status and β_{10} is the mean rate of teacher change. The level 2 models are below. π_{0i} is an indicator of variation in initial status and π_{1i} is an indicator of variation of growth.

$$\pi_{0i} = \beta_{00} + r_{0i}$$

$$\pi_{1i} = \beta_{10} + r_{1i}$$

It is possible that certain disruptions, such as changing schools or grades, could mask growth in beginning teachers' practices; therefore, using growth curve analysis allows me to control for these factors. A binary variable was created to indicate if a grade switch or school change had occurred. If the change occurred after the first year of teaching, the variable was equal to one in the second and third years of teaching to indicate the change from their initial placement. If the change occurred after the second year of teaching, the variable was equal to one in the third year of teaching. Two teachers switched grade levels in their second year of teaching and were switched back in their third year so their placement change variable was set back to zero in their third year.

Another change that could mask growth is year-to-year differences in the level of the observed class. Several teachers in our study taught both advanced and regular classes. An advanced class was defined based on the course title indicating that it was either an honors or advanced class. Due to scheduling issues, an advanced class might have been observed in one year and a regular class in another year. Whether an advanced or regular class was observed may impact a teachers' observed quality of teaching. For example, it may appear that the teacher did not improve their instructional quality if a teacher's advanced class was observed in their first year and a regular class was observed in the second year if the teacher lowered the rigor of their instruction for the regular class. Therefore, a binary variable for whether the observed class was honors or advanced was also included. Course data was gathered from teachers' schedules as well as observation notes completed by the videographers. The equation below shows the addition of the time-varying grade/school switch and advanced class indicators in level 1 of the model.

$$Y_{it} = \pi_{0i} + \pi_{1i}(Month\ of\ Observation)_{1i} \\ + \pi_{2it}(Grade\ or\ School\ Switch)_{2it} + \pi_{3it}(Advanced\ Class\ Observed)_{3it} + e_{it}$$

Relationship between Classroom Environment and Instructional Teaching Practices. Following the analysis of the development of beginning teachers' practices, I examined the relationship between beginning teachers' practices, as measured by the CLASS-S and math-specific IQA. To assess whether teachers who demonstrated strength in their classroom environment practices were also strong in their cross-subject and content-specific instructional practices at a particular time point, I used descriptive and correlational analysis. I estimated correlations at each time period for the CLASS-S domains of Emotional Support, Classroom

Organization, and Instructional Support; the individual IQA rubrics; and the Overall IQA score. The correlations provide initial evidence of the relationships between ratings of teachers' classroom environment and instructional teaching practices.

To obtain more information about the practical significance, or magnitude, of the relationships between ratings of teachers' classroom environment and instructional teaching practices. I examined means of CLASS-S scores for (a) teachers with averages of three or greater for each of the IQA Task Potential and Task Implementation rubrics and (b) teachers with averages below three for the Task Potential and Task Implementation rubrics. This illustrates differences in teachers' CLASS-S ratings on average if they assigned and implemented rigorous math instructional.

To further address the question of whether beginning teachers who demonstrate strength in their classroom environment practices are also strong in their cross-subject and content-specific instructional practices, I used scatter plots to examine the distribution of teachers' ratings over time. While the correlational analysis and the descriptive analysis described above assess the magnitude of the relationship of varying practices, they do not give a picture of the number of teachers (a) strong in both classroom environment and instructional practices, (b) strong at classroom environment practices and weaker at math-specific instructional practices, and (c) strong at math-specific instructional practices and weaker at classroom environment practices. The scatterplots provide a visual of the number of teachers in each category at each time period.

Next, I used hierarchical linear modeling with observations of practice nested within teachers to investigate whether reaching a particular level of competency on classroom environment practices, such as behavior and time management, made it more likely that a teacher implemented high-level cross-subject and content-specific instructional practices. I examined whether attaining “high” scores on Classroom Organization or Emotional Support was significantly associated with higher quality instructional practices. I used the Classroom Organization domain rather than the classroom management dimension score because the domain score captures aspects of productivity and instructional learning formats that are considered important for successful classroom management. Likewise, the Emotional Support domain scores capture a variety of cultural and socio-emotional teaching practices.

Binary indicators were used to capture high quality instructional practices. The binary indicator for high quality cross-subject instructional practices was equal to one if the teacher scored an average of five or higher on the Instructional Support domain of the CLASS-S. While the CLASS-S manual defines a 6 or 7 as “high” scores, a score of 5 was selected as the cutoff because (a) few teachers received a 6 or 7 limiting the variation of teachers in the “high” category and (b) teachers who received a 5 were on the cusp of the “high” category and did demonstrate better than average teaching practices. For math-specific instructional practices, the binary indicator was equal to one if the teacher scored an average of three or higher on the IQA Task Potential and Task Implementation rubrics. As stated previously, IQA developers note that a score of 3 or 4 on the IQA indicates high-level math instruction. To capture different levels of mastery of classroom environment levels (measured by the CLASS-S Emotional Support and Classroom Organization), a series of binary variables were used indicating whether teachers scored an average of 4 (“mid” level), 5 (high “mid” level), or 6 or 7 (“high” level).

To study the development of beginning teachers’ practice over their first three years of teaching, I used an average rating across the two days to represent practice at that time point. For this analysis, rather than using the average scores across the two days within each of the four time periods, each time period represented a unique observation on a particular day. Thus, some teachers had data from a total of eight days. Using the ratings gathered from the same observation day allows me to analyze the concurrent link of classroom environment and instructional teaching practices. The results show whether on a given day, a teachers' environment practices predict their attainment of a high level of instructional practice. Data were pooled across time periods to examine the association between environment and instructional practices across the beginning teachers' first three years.

The equation shown below is an extension of the growth model used in the prior analysis. Here, the outcome is the binary indicator of "high" instructional practices. The model predicts a teacher's instructional practices on a particular day *d* as a function of their classroom environment practices on the same day, controlling for other factors that may influence teacher's practices, such as whether they were teaching an honors or advanced class, using a reform math curriculum, and school FRPL percentage.

$$IP_{id} = \pi_{0i} + \pi_{1id}(Month\ of\ Observation)_{1id}$$
$$+ \pi_{2id}(Classroom\ Environment\ Practices)_{2id} + \pi_{3it}(Control\ Variables)_{3id} + e_{id}$$

As a sensitivity test, I used a teacher fixed effects approach that allows me to compare observations at different times within teacher. The approach holds constant factors that are the same within a teacher over time and only includes covariates that change over time. For example, the variable indicating reform math curriculum is omitted because the curriculum is the same over time within teachers. Y_{id} represents an individual teacher's instructional practice rating on a particular day. The instructional practice rating is a function of the teacher's classroom environment rating from the same day, whether the teacher was observed teaching an advanced or honors math class, school free and reduced price lunch during the year of the observation, and the teacher's fixed characteristics δ_i. Only teachers with variation in their measure of instructional practices are included in the fixed effect analysis.

$$IP_{id} = \beta_0 + \gamma_1 Classroom\ Environment\ Practices_{it} + \gamma_2 Control\ Variables_{it} + \delta_i + \varepsilon_{it}$$

It is important to note that it is difficult to disentangle the relationship of these different aspects of teaching practice because these practices may influence each other simultaneously. For example, reaching proficiency on management practices may enable teachers to implement more challenging instructional practices that allow for less teacher-directed work. However, more engaging instructional practices may also lead to better classroom management. While time order of the association between the practices is difficult to establish, I can investigate whether attaining a "high" level classroom environment practices in a prior time period is associated with reaching a "high" level of instructional practices by included lagged measures of classroom environment practices rather than concurrent measures. HLM models are used to estimate the predicated probability of having "high" instructional practices based on whether the teacher achievement a "high" average rating of classroom environment practices from the prior time period. For example, Classroom Organization scores from spring of year 1 were used to predict whether a teacher implemented rigorous math instruction on observation day 1 of spring of year 2, as well as on observation day 2 of spring of year 2. In these models, I focus specifically on whether teacher reached a "high" level of environment practices, rather than using several indicators of environment practice levels. Therefore, the results from this analysis show if reaching a "high" level of environment practices means a teacher is likely to have "high" instructional practices in the following time period. Like the other methods described in this section, the results are intended to foster a better understanding of the relationship between teachers' environment and instructional practices.

Association between Teaching Practices and Teacher Effectiveness at Increasing Student Achievement. To assess the relationship between evaluations of teaching practices and value-added estimates of teacher effectiveness, I used correlational analysis, which is the method most frequently used by studies assessing this relationship. Correlations between teachers' ratings on the CLASS and the IQA with their value-added estimates provide an indicator of the relative strength of two teacher quality indicators.

I also examined whether there are differences in the teaching practices of teachers with higher and lower value-added scores. For this analysis, teachers' value-added estimates were used to place them in quartiles for each time period. I examined whether teachers who are in the top quartiles of value-added scores have higher ratings of teaching practices compared to teachers in the lower quartiles. For each time period, I tested the statistical significance of the rating differences across the value-added quartile groups (e.g., Grossman et al., 2010). This method helps us understand whether "teachers who tend to promote higher student achievement growth are teaching differently than teachers associated with lower student achievement growth" (Kane et al., 2010, p. 18).

The correlational analysis and quartile comparisons provide useful information regarding the relationship between ratings of teacher practice and value-added scores and allow me to compare my results to other studies. While most other studies use correlational analysis and compare ratings across quartile indicators of value-added scores, a benefit of this study is the multiple observations over time for individual teachers. Thus, I used hierarchical linear modeling with value-added scores and observation ratings nested within teacher to further investigate their association.

For each area of instructional practice, I ran a separate model to examine the relationship between that practice and teachers' value-added scores. I also tested whether the results held controlling for other practice ratings. I did not include Instructional Support and Overall IQA ratings in the same model because they capture many of the same aspects of practice. Including both the measures of cross-subject and math-specific practices in the same model would make it difficult to distinguish the relationships because of their conceptual overlap. Since we would expect beginning teachers to improve their value-added over time, I included binary indicators for years 2 and 3. I also included interactions between the practice scores and year to test if the strength of relationships differs in different time periods. All models controlled for whether the teacher experienced a grade or school change. As the basic model below shows, teacher value-added is a function of teacher practice, year, and control variables such as other practices and a grade/school change indicator.

$$VA_{it} = \pi_{0i} + \pi_{1it}(Teacher\ Practice)_{1it}$$
$$+ \pi_{2it}(Year)_{2it} + \pi_{3it}(Control\ Variables)_{3it} + e_{it}$$

I also used a teacher fixed effects analysis to examine the relationship between ratings of teacher practice and value-added scores. In these models, the relationship between value-added scores and practice ratings are examined within teachers. This method holds constant time-invariant teacher characteristics.

The final analysis addresses the question of whether change in teachers' practice is associated with change in their value-added scores. Change scores were calculated by subtracting the practice ratings or value-added scores from the previous year's score. In this model, one-year change scores are substituted in place of ratings. Two change scores (year 1 to year 2 and year 2 to year 3) were calculated based on the three years of data. These models controlled for change in other practices and grade/school change. Of note is that this analysis only includes teachers with change scores, meaning they had to have data in both spring of year 1 and year 2 or in year 2 and year 3. Thus, the sample size is reduced and power for detecting significant relationships is also diminished.

$$Change\ in\ VA_{it} = \pi_{0i} + \pi_{1it}(Change\ in\ Teacher\ Practice)_{1it}$$

$$+ \pi_{2it}(Control\ Variables)_{3it} + e_{it}$$

CHAPTER IV

RESULTS

In this section, I first present findings regarding beginning middle school math teachers' initial levels of classroom environment and instructional practices and the extent to which they improve on various aspects of their practices during their first three years of teaching. Next, I report findings from an investigation of the relationship between teachers' environment and instructional practices. Finally, I describe the outcomes of the analysis investigating the association between beginning teachers' classroom environment and instructional teaching practices with their effectiveness at increasing student learning.

Development of Beginning Teachers' Practices

Research finds that beginning teachers struggle with both classroom environment and instructional practices, but few studies have examined the initial quality of beginning teachers' practices or which types they are more likely to improve. I find that at the beginning of their careers, the middle school math teachers in this study rated higher on their classroom environment practices than on their instructional practices. Figure 1 shows the distribution of teachers rated as low, mid, and high on the CLASS-S and math-specific IQA in the fall of year 1. Only two teachers rated "low" in Classroom Organization and none rated "low" in Emotional Support, whereas 7 rated "low" in Instructional Support. On the IQA rubrics, no teachers rated "low" on Task Potential and Task Implementation, but over 60% (n=38) rated "low" on Discussion. Since these three aspects of practice are combined to represent math-specific instructional practice, the lower discussion scores bring down the overall ratings of math-specific instructional practice.

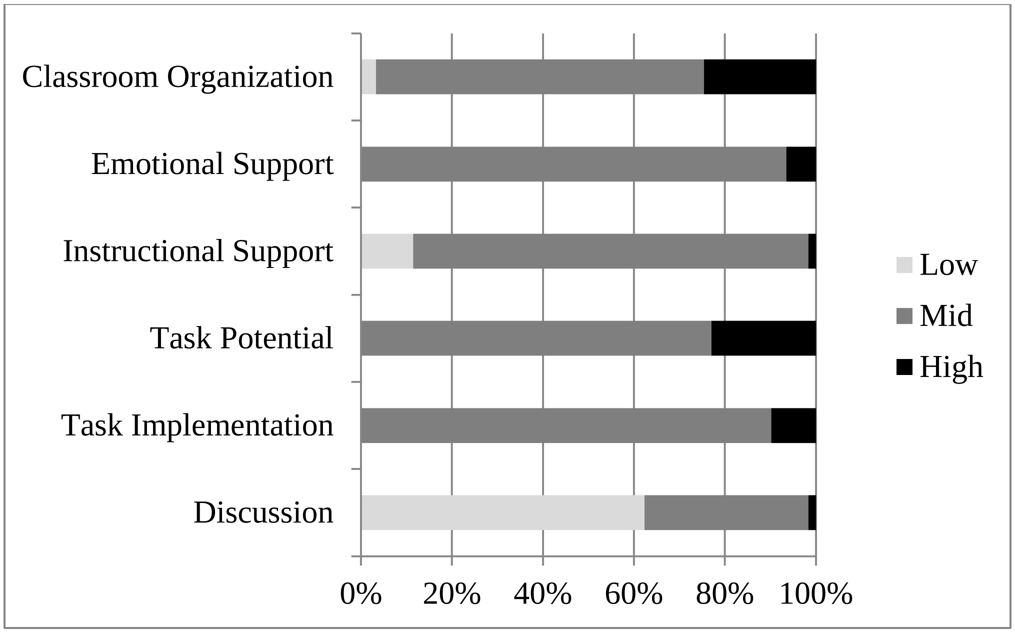


Figure 1. Percentage of Teachers Rated as Low, Mid, or High on their Fall Year 1 Teaching Practices

The middle school math teachers continued to have higher scores for classroom environment practices than instructional practices across their first three years of teaching. Average scores for classroom environment and cross-subject instructional practice ratings increased over time (see Figure 2). The CLASS-S ratings were not significantly different from the fall to spring of year 1, but all three dimensions were significantly greater from the fall and spring of year 1 to the spring of year 2 (Emotional Support: $p<0.01$; Classroom Organization: $p<0.001$; Instructional Support: $p<0.05$) and to the spring of year 3 (Emotional Support and Classroom Organization: $p<0.001$; Instructional Support: $p<0.01$). Increases in the CLASS-S domain averages were about one-fourth to half a point from year 1 to year 2 and one-tenth to one-sixth of a point from year 2 to year 3. Still, in all four time periods the average scores for all CLASS-S domains were in the “mid” category, with the Classroom Organization and Emotional Support averages nearer to the “high” category and the Instructional Support average nearer to the “low” category.

A score in the “mid” range on the CLASS-S indicates that some of the dimension indicators were present during the observation. The teachers’ average of 4.47 in fall of year 1 on Emotional Support indicates some evidence of positive climate, teacher sensitivity and regard for adolescent perspective and occasional evidence of a negative climate, while an their average of 4.71 for Classroom Organization indicates some evidence of behavior expectations, routines, planning, and monitoring, though these may not always be effective (1 to 7 scale). The 3.54 average on Instructional Support indicates limited content understanding practices and questioning strategies with occasional opportunity for higher-order thinking through analysis problem solving tasks. Increases in CLASS-S ratings mean that the indicators were present more frequently. For example, a teacher’s rating could move from a 4 to a 4.5 on Classroom Organization if the teacher demonstrated more frequent proactive monitoring of student behavior, along with more frequently maximizing learning time.

For math-specific instructional practices, teachers at the beginning of their first year assigned mostly procedural tasks and had limited classroom discussion of math concepts. In the fall of year 1, teachers had an average Task Potential score of 2.33 and a Task Implementation average of 2.09 (0 to 4 scale). The lower scores for implementation indicate that some teachers did not implement the tasks to their potential for engaging students in rigorous math activity. On average, teachers scored lowest on the Discussion rubric, indicating student responses characterized by one-word answers. These trends were maintained across all four time periods, and teachers’ average scores on the IQA rubrics did not change much over time (see Figure 2). No significant differences were found when comparing teachers' Overall IQA scores in consecutive years, nor were significant difference found when comparing teachers’ scores from year 1 to year 3.

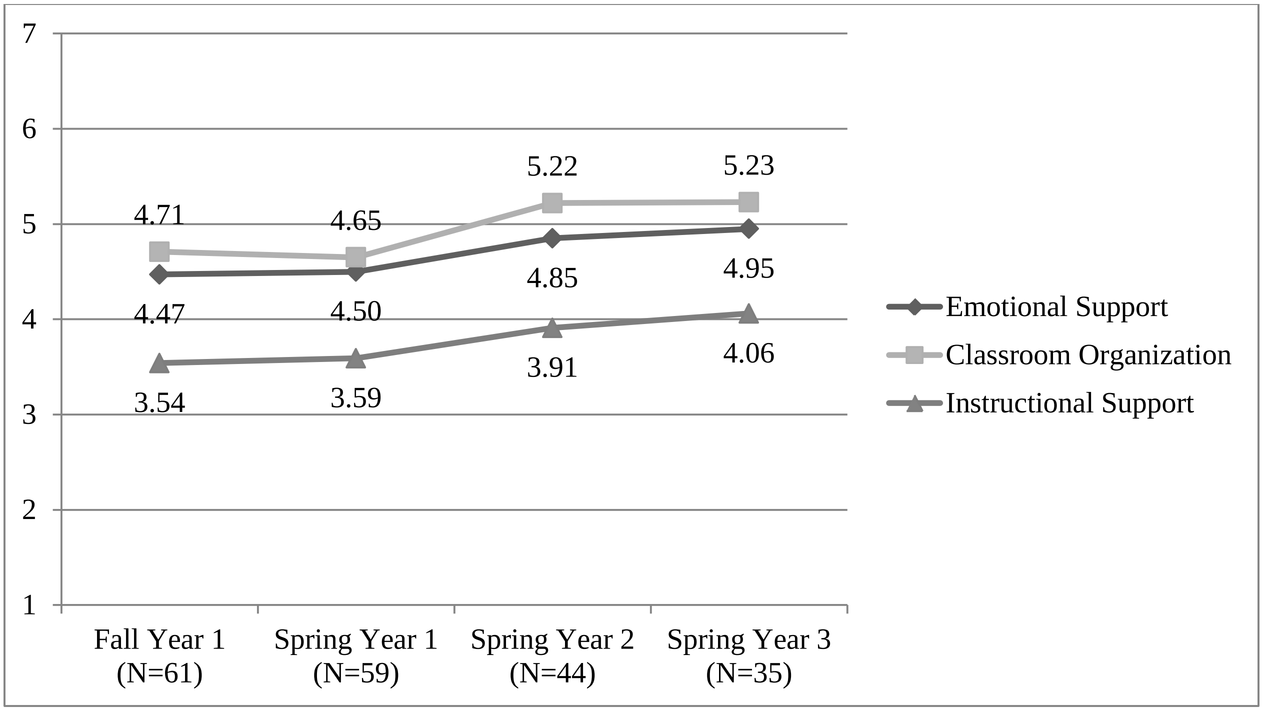


Figure 2. CLASS-S Scores over Time

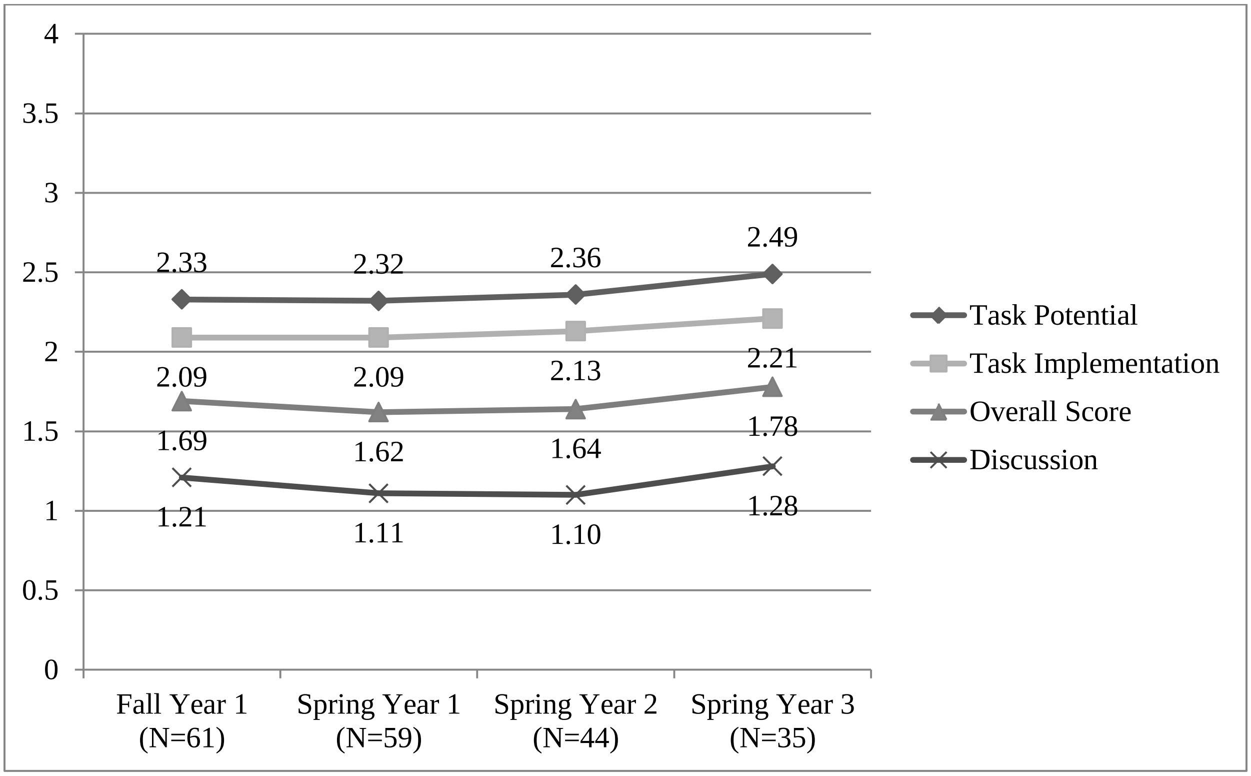


Figure 3. Math-specific IQA Scores over Time

While mean scores show how the average teacher performed, it is also useful to examine the range of ratings of teaching practices. Table 11 provides the standard deviations and minimum and maximum scores for CLASS-S and IQA ratings. Of the CLASS-S domains, the widest variation among teachers' scores was in the Classroom Organization domain. This indicates that some teachers entered the profession with "high" classroom management skills, whereas others had "low" management practices, rather than all being about the same. Figure 1 provides an illustration of this variation in the fall of year 1, with 25% of teachers rating "high," 72% rating "mid," and 3% rating "low." In contrast, for Emotional Support 93% of teachers rated in the "mid" category in fall of year 1. The variation of Classroom Organization ratings significantly decreased over time, which was likely due to the "low" management teachers improving their skills or exiting the sample. The standard deviations for Classroom Organization ratings were significantly different from year 1 to year 3 ($p<0.05$). Standard deviations for Emotional Support and Instructional Support were not significantly different over time.

Of the IQA rubrics, variation among teachers was widest for Discussion and smallest for Task Implementation, though none of the differences in variation across time are statistically significant. As shown in Figure 1, 90% of the beginning teachers rated in the "mid" category for task implementation in fall of year 1, whereas there was more variation in the quality of the discussion in the beginning teachers' classrooms.

The variation of CLASS-S and IQA ratings was slightly smaller for this sample of beginning middle school math teachers than has been reported by other studies using the CLASS-S and IQA to rate teaching practices of teachers across subjects and experience levels (Bell et al., 2012; Grossman et al., 2010; Matsumura et al., 2006).[7] This indicates that there is

[7] Bell et al. (2012) reported somewhat larger CLASS-S rating standard deviations of 0.95 for Emotional Support, 0.84 for Classroom Organization, and 0.98 for Instructional Support. These are about one to two-tenths higher than

less variation in teaching practices of beginning teachers from the same subject area than there is for teachers across subjects and experience levels.

Table 11. CLASS-S and IQA Mean, Minimums, and Maximums over Time

	Fall Year 1 (N=61)			Spring Year 1 (N=59)			Spring Year 2 (N=44)			Spring Year 3 (N=35)		
	Mean (SD)	Min	Max	Mean (SD)	Min	Max	Mean (SD)	Min	Max	Mean (SD)	Min	Max
CLASS-S (1-7 scale)												
Emotional Support	4.47 (.74)	2.56	5.56	4.50 (.81)	2.94	6.00	4.85 (.74)	3.13	6.50	4.95 (.70)	3.25	6.19
Classroom Organization	4.71 (1.06)	2.25	6.33	4.65 (1.12)	2.00	6.50	5.22 (.87)	2.78	6.67	5.23 (.75)	3.22	6.50
Instructional Support	3.54 (.76)	1.63	5.94	3.59 (.79)	1.81	5.69	3.91 (.69)	2.65	5.25	4.06 (.75)	2.69	5.81
Math-specific IQA (0-4 scale)												
Task Potential	2.33 (.57)	1.50	4.00	2.32 (.55)	1.50	3.75	2.36 (.63)	1.00	4.00	2.49 (.67)	1.25	4.00
Task Implementation	2.09 (.40)	1.50	3.25	2.09 (.40)	1.50	3.50	2.13 (.48)	1.00	3.50	2.21 (.52)	1.25	3.75
Discussion	1.21 (.67)	0.00	2.75	1.11 (.72)	0.00	2.25	1.10 (.70)	0.00	2.25	1.28 (0.81)	0.00	3.00
Overall IQA	1.69 (.46)	0.75	3.15	1.62 (.38)	0.88	2.31	1.64 (.40)	.71	2.60	1.78 (.49)	1.00	3.13

Using growth curve analysis, I found that over their first three years of teaching, beginning math teachers experienced statistically significant improvement in the areas of classroom environment practices and cross-subject instructional practices but not in math-specific instructional practices.[8] Table 12 presents the growth parameters from the HLM models for each of the CLASS-S domains and the Overall IQA scores, and Table A2 in the Appendix provides more detailed results. The coefficient on the variable month was significant for all three CLASS-S domains, but not for the Overall IQA score, indicating no significant improvement on

the standard deviations for CLASS-S scores in year 3. Grossman et al. (2010) reported standard deviations at the dimension level of the CLASS-S, ranging from 1.08 for regard for adolescent perspective (Emotional Support) and 1.60 for Behavior Management (Classroom Organization). Matsumura et al. (2006) reported standard deviations of 0.91 for Task Potential, 0.74 for Task Implementation, and 1.38 for Discussion.

[8] As a reminder, the Overall IQA was used as a continuous indicator of math-specific instructional practices.

the IQA over their first three years of teaching. The growth curve models controlled for change in grade or school in a particular year and whether the course observed was honors or advanced.

Teachers were predicted to improve most on their Classroom Organization ratings – by about a third of a point in a year and almost a point over three years. On their Emotional Support and Instructional Support ratings, teachers were predicted to increase by about a fifth of a point over a year. Over three years, this was an increase of about two-thirds of a point. An increase of a third of a point was about a third of a standard deviation for Classroom Organization, and an increase of one-fifth of a point was about one-fourth of standard deviation for Emotional Support and Instructional Support.

Table 12. Growth Coefficients[9] Indicating Monthly Improvement in Teaching Practice (N=199)

	Overall IQA	CLASS-S Emotional Support	CLASS-S Classroom Organization	CLASS-S Instructional Support
Month	0.002	.019***	0.025***	0.019***
	(0.00)	(0.00)	(0.01)	(0.00)
Constant	1.58***	4.33***	4.50***	3.35***
	(0.06)	(0.11)	(0.16)	(0.11)

+ for p<.10, * for p<.05, ** for p<.01, and *** for p<.001

One concern is that the improvement in beginning teachers' practices was due to teachers with lower practice ratings leaving the profession. However, when including only teachers who remained in the study all three years similar results were found, indicating that teachers did improve. The only slight difference was that the growth parameter month for Instructional Support was slightly smaller (see Table A3 in Appendix). This suggests that teachers with lower quality cross-subject instructional practices were more likely to exit. Another study using AIM

[9] The coefficient on month indicates improvement on the outcome measure during one month. Coefficients were multiplied by 12 to indicate improvement in one year and by 36 to indicate improvement over three years.

data found no significant differences in the IQA ratings of teachers who remained and those who either moved schools or left the profession after their first year. They found significant differences on spring CLASS-S Emotional Support and Instructional Support scores between stayers and movers/leavers (Neergaard, Dunn, Smith, & Desimone, 2012).

Beginning middle school math teachers started teaching with higher ratings of classroom environment practices and were more likely to improve their classroom environment practices during their first three years in the classroom. While the beginning teachers started in the low end of the “mid” level range on measures of both cross-subject and math-specific instructional practices, they improved their cross-subject instructional practices, but not their math-specific instructional practices. The next section explores the relationship between measures of teachers’ classroom environment and instructional teaching practices in an effort to understand whether development of some practices was related to growth in others.

Relationship between Classroom Environment and Instructional Teaching Practices

While it is helpful to better understand which areas of practice they may be more likely to improve, it is also important to consider the development of certain practices in relation to others. I examined the correlations among ratings of the various practices to see whether teachers who performed well at certain practices were also likely to have high ratings of other areas of practice. Table 13 shows the correlations between teachers’ Overall IQA scores and their CLASS-S domain scores at each time period. Overall IQA scores are correlated most highly with the Instructional Support domain ratings across all time periods. The higher correlation with Instructional Support is expected since the IQA and Instructional Support domain are both focused on aspects of instructional practices, whereas the other two domains are focused on classroom environment practices. Interestingly, the strength of the relationship between the

Overall IQA ratings and Classroom Organization and Emotional Support ratings declined over time. One explanation for this is that teachers are improving in their classroom environment practices and not in their math-specific instructional practices which weakens the relationship over time.

Of the individual IQA rubric scores, Task Implementation was generally most highly correlated with Instructional Support (see Table 14). Task Implementation and Discussion were noted to be similar to several of the indicators found under Instructional Support so the moderately-sized correlations with these rubric ratings makes sense as they capture some of the same aspects of instructional practice (see Table 10). For example, Task Implementation captures similar practices to the analysis and problem solving dimension, such as whether students engaged in higher-order thinking, and Discussion rates teachers on the depth of conversation following work time, which is similar to practices captured in the quality of feedback and instructional dialogue dimensions of the CLASS-S. The correlations between the CLASS-S domain scores over time are significant and generally high amongst all the domain scores but are strongest between Emotional Support and Classroom Organization, which makes sense as they both assess classroom environment practices (see Table 15).

Table 13. Correlations between Overall IQA Scores and CLASS-S Scores over Time

	Math-specific Overall IQA Scores			
CLASS-S Domain Scores	Fall Year 1 (N=61)	Spring Year 1 (N=59)	Spring Year 2 (N=44)	Spring Year 3 (N=35)
Emotional Support	0.40**	0.42***	0.23	0.19
Classroom Organization	0.35**	0.38**	0.29+	0.12
Instructional Support	0.54***	0.53***	0.33*	0.52**

+ for p<.10, * for p<.05, ** for p<.01, and *** for p<.001

Table 14. Correlations between IQA Rubric Scores and CLASS-S Instructional Support Scores over Time

	CLASS-S Instructional Support Scores			
IQA Rubric Scores	Fall Year 1 (N=61)	Spring Year 1 (N=59)	Spring Year 2 (N=44)	Spring Year 3 (N=35)
Task Potential	0.49***	0.22+	0.26+	0.47**
Task Implementation	0.52***	0.29*	0.39**	0.51**
Discussion	0.45***	0.47***	0.20	0.40*

+ for p<.10, * for p<.05, ** for p<.01, and *** for p<.001

Table 15. Correlations between CLASS-S Domain Scores over Time

	Fall Year 1 (N=61)	Spring Year 1 (N=59)	Spring Year 2 (N=44)	Spring Year 3 (N=35)
Emotional Support & Classroom Organization	0.83***	0.79***	0.84***	0.77***
Emotional Support & Instructional Support	0.79***	0.80***	0.84***	0.72***
Classroom Organization & Instructional Support	0.73***	0.74***	0.72***	0.71***

+ for p<.10, * for p<.05, ** for p<.01, and *** for p<.001

Further descriptive analysis demonstrates the magnitude of the relationships among the ratings of teaching practices. Figures 4 and 5 illustrate how much higher teachers CLASS-S ratings are on average if they assigned and implemented “high” math instruction, as measured by the IQA. As a reminder, “high” math instruction, as classified by the IQA, occurs when teachers assign and implement open-ended tasks and facilitate discussion that requires students to provide evidence of their reasoning and understanding and connect mathematical ideas. Rubric scores of 3 or 4 indicate “high” math instruction.

As expected, given the positive correlations between CLASS-S scores and Overall IQA scores, in classrooms where rigorous math activity was present, teachers were more frequent users of the practices measured by the CLASS-S, such as making connections to current life and using a variety of instructional learning formats. Like the correlations, the strength of the relationship varied across domains and time periods. Overall, teachers who assigned rigorous math tasks scored an average of about one-third of a point higher on the CLASS-S domain scores (see Figure 3), and teachers who implemented rigorous math activities scored an average of a half point to two-thirds of a point higher on the CLASS-S domain scores (see Figure 4). I found larger differences in average CLASS-S Instructional Support ratings between teachers with rigorous and non-rigorous implementation than the differences in the other CLASS-S domains. This finding reflects the stronger correlations found between the Overall IQA and Instructional Support scores.

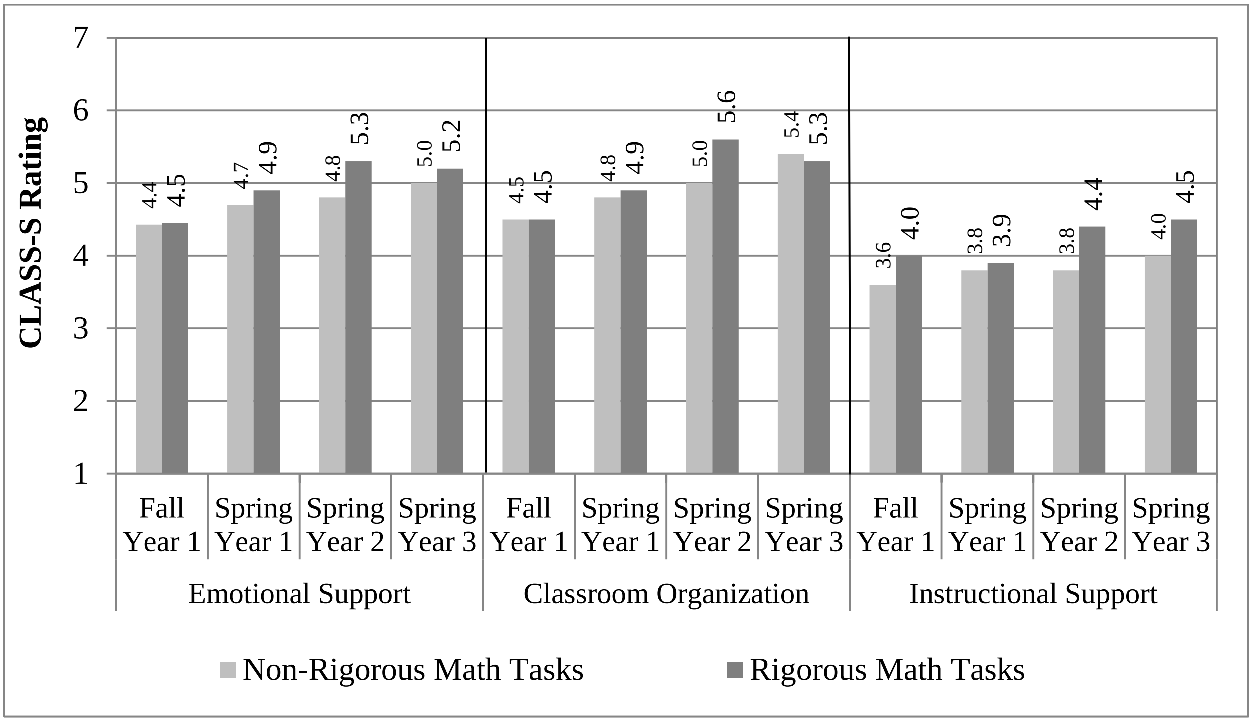


Figure 4. CLASS-S Domain Means by Rigorous Math Implementation over Time

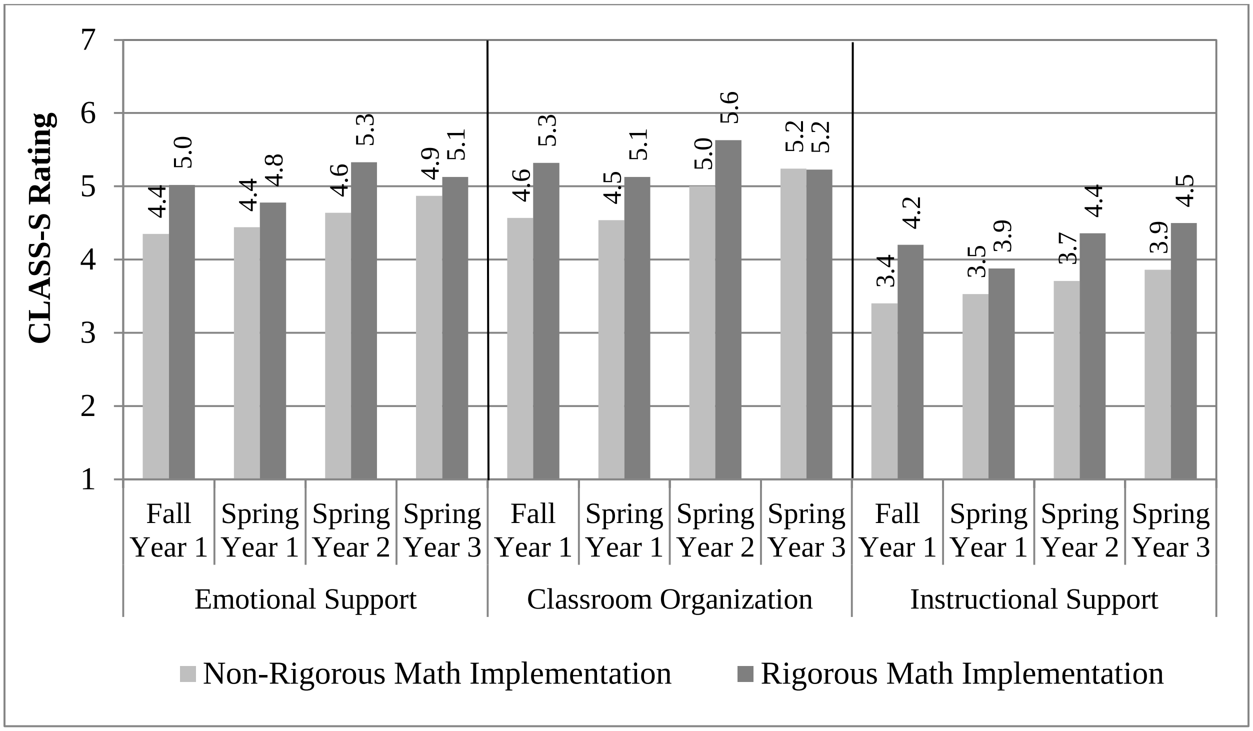
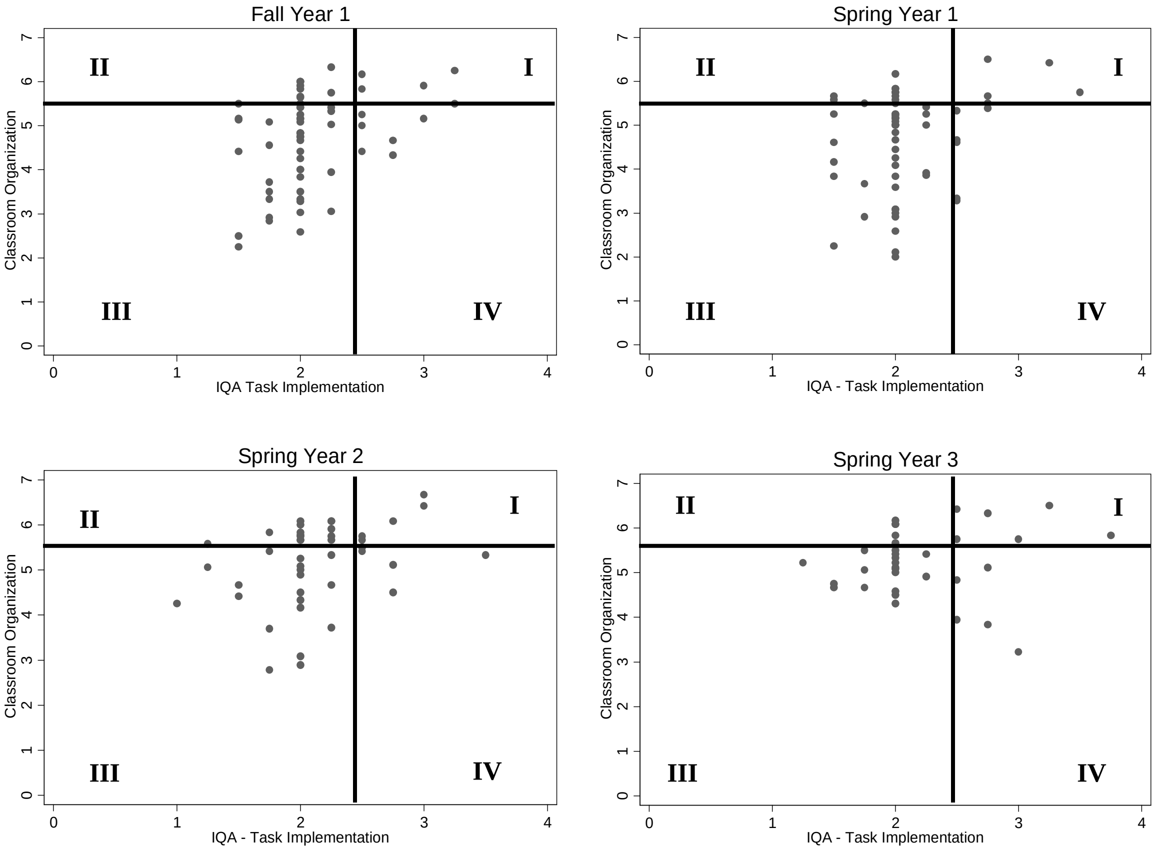


Figure 5. CLASS-S Domain Means by Rigorous Math Implementation over Time

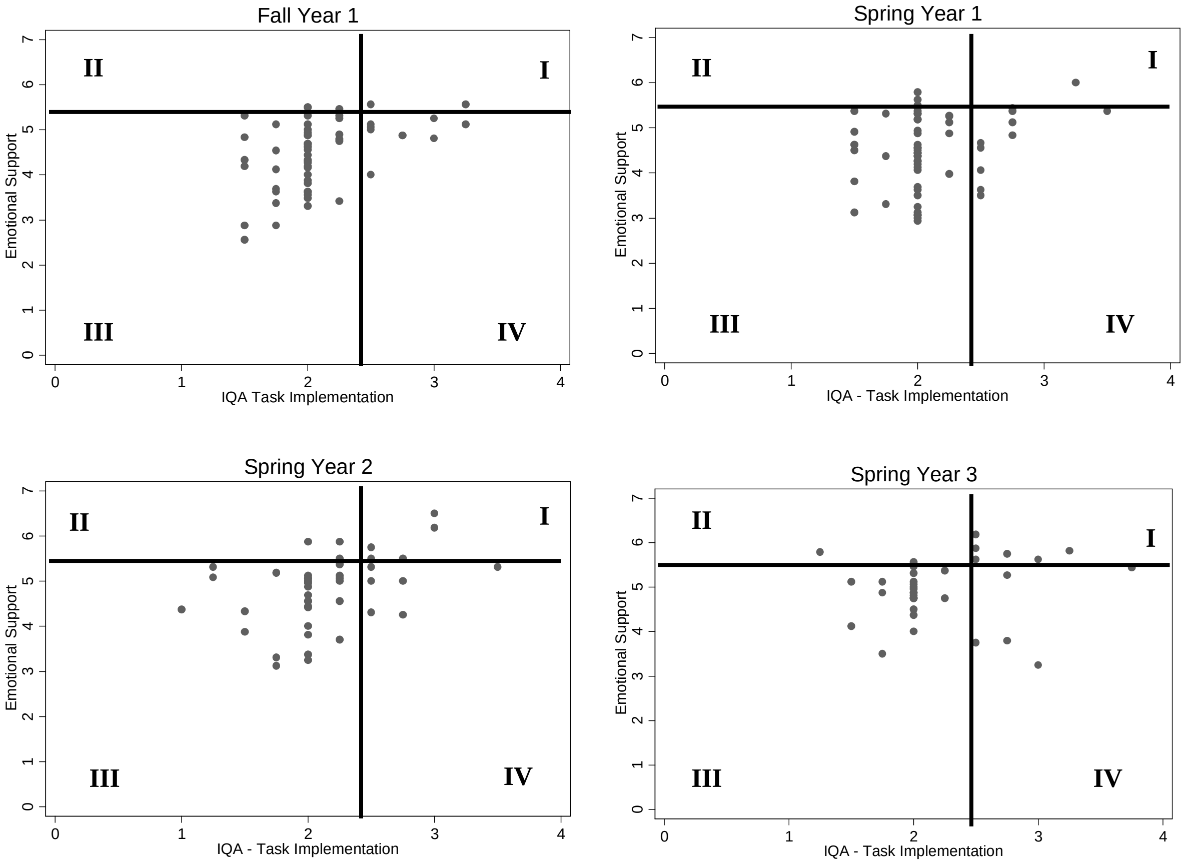
While the correlations and figures above illustrate that teachers implementing rigorous math activity also have higher levels of classroom environment and cross-subject instructional practices, further descriptive analysis reveals that that (a) some teachers are strong in both areas of practices, (b) some teachers were strong at classroom environment practices and weaker at math-specific instructional practices, (c) some teachers were strong at math-specific instructional practices and weaker at classroom environment practices, and (d) some teachers were weak in both areas of practice. Figures 6 and 7 are scatterplots of teacher's scores on the CLASS-S and IQA illustrate the distribution of teachers' ratings across both classroom environment and instructional practices. In the plots, I focus on the relationship between implementing rigorous math activity (Task Implementation scores) and classroom environment practices (Emotional Support and Classroom Organization scores). Lines are placed at cutoffs for what is considered a "high" level of practice (i.e., an average of 6 or higher on the CLASS-S and an average of 3 or higher on the IQA).

Few beginning teachers demonstrated strength in both their classroom environment and math-specific practices, and it was more common for teachers to rate highly on classroom environment practices and low on math-specific instructional practices than it was for them be strong at the implementing rigorous math instruction and weak on classroom environment practices. The number of teachers rated at the "high" level on both Task Implementation and Classroom Organization (8 to 17%) and both Task Implementation and Emotional Support (2 to 17%) increased over time. Still, many teachers (37-43% for Classroom Organization; 63-79% for Emotional Support) were not in either "high" level of classroom environment or math-specific instructional practices in each time period. Across all four time periods, only one teacher who scored "high" on Task Implementation scored below a 4 on Classroom Organization, which indicates “mid” level practices. Yet, many teachers who rated highly on Classroom Organization or Emotional Support scored an average of 2 or below on Task Implementation.



Quadrant I: Strong in Classroom Organization and Task Implementation
Quadrant II: Strong at Classroom Organization, Weaker at Task Implementation
Quadrant III: Weaker in Classroom Organization and Task Implementation
Quadrant IV: Strong in Task Implementation, Weaker in Classroom Organization

Figure 6. Scatterplot of IQA Task Implementation and CLASS-S Classroom Organization Scores



Quadrant I: Strong in Classroom Organization and Task Implementation
Quadrant II: Strong at Classroom Organization, Weaker at Task Implementation
Quadrant III: Weaker in Classroom Organization and Task Implementation
Quadrant IV: Strong in Task Implementation, Weaker in Classroom Organization

Figure 7. Scatterplot of IQA Task Implementation and CLASS-S Emotional Support Scores

In my analysis of whether teachers reaching a particular level of competency on classroom environment practices were more likely to implement high-level instructional practices, I find that beginning teachers who scored 5 or higher on Emotional Support and Classroom Organization were significantly more likely to both assign and implement rigorous math instruction. Table 16 shows the predicted probability of “high” instructional practices as a function of their Classroom Organization and Emotional Support scores for three outcomes: IQA Task Potential, IQA Task Implementation, and CLASS-S Instructional Support scores. For example, scoring a 5 average on Classroom Organization was associated with being 1.2 times more likely to score a 3 or 4 on Task Implementation and Task Potential compared to scoring below a 4 average, and scoring a 6 or higher was associated with being 1.4 times as likely to rate highly on Task Potential and 1.6 times as likely to rate highly on Task Implementation. This provides evidence of a statistically significant relationship between mastery of classroom environment practices and ability to implement rigorous math instruction at a particular time period.

Table 16. Predicted Probability of "High" Instructional Practices Given Classroom Environment Practice Scores from the Same Observation Period (HLM)

	High Task Potential (3 or 4)		High Task Implementation (3 or 4)		High Instructional Support (5, 6, or 7)[10]
	Classroom Organization	Emotional Support	Classroom Organization	Emotional Support	Classroom Organization
Level 4 Average	0.55 (0.51)	0.62 (0.52)	0.71 (0.64)	0.80 (0.72)	NA
Level 5 Average	1.25* (0.49)	1.59** (0.52)	1.21* (0.61)	1.91** (0.71)	NA
Level 6 or 7 Average	1.43** (0.49)	1.94** (0.60)	1.58** (0.61)	2.64*** (0.76)	14.54*** (6.77)
Controls	X	X	X	X	X
Observations	387	387	387	387	387

+ for p<.10, * for p<.05, ** for p<.01, and *** for p<.001

My findings from a similar analysis using teacher fixed effect models also show that attaining a score of 5 or 6 on Emotional Support was significantly associated with the likelihood of having "high" math-specific and cross-subject instructional practices, but the relationship

[10] Due to limited variation in Emotional Support/Classroom Organization and Instructional Support scores, I was unable to analyze the relationship of concurrent classroom environment and cross-subject instructional practices using a variety of levels of Emotional Support/Classroom Organization cutoffs. Of 69 observations where teachers scored a 5 average or higher on Instructional Support, only one observation had a score below a 5 average on Emotional Support and Classroom Organization. Furthermore, as the same rater scored a teacher on Emotional Support, Classroom Organization, and Instructional Support during the same observation day, there are some concerns about rater bias when comparing these practice ratings from the same observation.

between higher levels of Classroom Organization practices and higher levels of math-specific instructional practices is not statistically significant (see Table 18). The magnitudes of the predicted probabilities were similar to findings from the HLM models for the relationship between Emotional Support and "high" math-specific instructional practices. I also found that teachers were 2.2 to 2.5 times more likely to implement "high" cross-subject instructional practices if they demonstrated "high" levels of classroom environment practices.

Table 17. Predicted Probability of "High" Instructional Practices Given Classroom Environment Practice Scores from the Same Observation Period (Fixed Effects)

	High Task Potential (3 or 4)		High Task Implementation (3 or 4)		High Instructional Support (5, 6, or 7)	
	Classroom Organization	Emotional Support	Classroom Organization	Emotional Support	Classroom Organization	Emotional Support
Level 4 Average	0.07	0.51	-0.13	0.73	NA	NA
	(0.56)	(0.62)	(0.63)	(0.83)		
Level 5 Average	1.02	1.62*	0.54	1.98*	NA	NA
	(0.55)	(0.64)	(0.60)	(0.87)		
Level 6 or 7	1.01	1.77*	0.67	2.35*	2.20***	2.47***
Average	(0.58)	(0.73)	(0.63)	(0.94)	(0.44)	(0.58)
Observations[11]	322	322	288	288	198	198

+ for p<.10, * for p<.05, ** for p<.01, and *** for p<.001

Demonstrating mastery of Emotional Support practices in a prior time period was also associated with assigning rigorous tasks, but no significant relationship was found for prior Classroom Organization ratings (see Table 18). No significant relationships were found between prior mastery of classroom management skills and future proficiency in instructional practices.

[11] Only teachers with variation in their measure of instructional practices are included in the fixed effect analysis. 13 teachers had no variation in the "high" Task Potential ratings, 17 teachers had no variation in their "high" Task Implementation ratings, and 34 teachers had no variation in the "high" Instructional Support ratings.

Table 18. Predicted Probability of “High” Instructional Practices Given Prior Classroom Environment Practice Scores from the Prior Observation Period (HLM)

	High Task Potential (3 or 4)		High Task Implementation (3 or 4)		High Instructional Support (5, 6, or 7)	
	Classroom Organization	Emotional Support	Classroom Organization	Emotional Support	Classroom Organization	Emotional Support
Level 6 or 7	-0.17	1.39*	-0.52	1.15	-0.06	-1.05
Average	(0.42)	(0.61)	(0.50)	(0.62)	(0.57)	(0.94)
Controls	X	X	X	X	X	X
Observations	265	265	265	265	261	261

+ for p<.10, * for p<.05, ** for p<.01, and *** for p<.001

In summary, demonstrating proficiency in some areas of practice often means that a teacher will be skilled at other areas of practice. This was especially true for practices that are conceptualized more similarly, like cultural/emotional practices and management practices or practices related to implementing instruction that promotes higher-order thinking skills. It was also true for practices with less conceptual overlap, like overall ratings of environment and instructional practices. As the correlational and descriptive analysis showed, on average teachers who had higher ratings of environment practices also had higher ratings of instructional practices. Still, some teachers performed well in some areas but poorly on others. It was more common for teachers to rate highly on classroom environment practices and low on math-specific instructional practices than it was for them be strong at the implementing rigorous math instruction and weak on classroom environment practices. The fixed effects analysis reinforced this finding for the relationship between cultural/emotional practices and “high” instructional practices. Reaching a “level 5” on the CLASS-S Emotional Support domain increased the chance that a teacher implemented high levels of instructional practices during the same lesson. Additionally, reaching a “level 6” was associated with an even greater probability of proficiency in instructional practices. In other words, the better a beginning teacher is at their environmental practices, the more likely they will be able to deliver high-level instruction.

Association between Teaching Practices and Teacher Effectiveness

While it appears that being good at some practices often means a teacher will be good at others, not all practices may be equal when it comes to increasing student achievement. Correlations between value-added scores and teaching practice ratings indicate that classroom environment practices are important for increasing student learning in the first two years of teaching, but instructional practices are the most highly correlated with teacher effectiveness ratings by the third year. Table 19 presents the correlations between value-added scores and teachers' practice ratings during each time period. While the correlations with value-added scores and Classroom Organization scores were highest in years 1 (r=0.24) and 2 (r=0.30), they were lowest in year 3 (r=0.04). Conversely, math-specific instructional practices, measured by Overall IQA scores, had a correlation of 0.52 with value-added scores in year 3, and cross-subject instructional practices, measured by Instructional Support, had a correlation of 0.34.

Table 19. Correlations between Teachers' Value-added Scores Practice Ratings

	Fall Year 1 (n=57)	Spring Year 1 (n=56)	Year 1 Average (n=55)	Spring Year 2 (n=42)	Spring Year 3 (n=28)	Across Time
Overall IQA	0.08	0.02	0.05	0.30^{+}	0.52**	0.26*
Emotional Support	0.16	0.08	0.14	0.30*	0.08	0.16^{+}
Classroom Organization	0.24^{+}	0.26^{+}	0.28*	0.30^{+}	0.04	0.22*
Instructional Support	0.21	0.05	0.16	0.28^{+}	0.34^{+}	0.21*
Overall CLASS-S	NA	NA	NA	NA	NA	0.22*

+ for p<.10, * for p<.05, ** for p<.01, and *** for p<.001

The correlations between value-added scores and ratings of teachers' practices for the sample of teachers who remained in the study all three years were mostly similar to the full sample. The main difference was a stronger and marginally significant correlation between value-added scores and Instructional Support ratings from fall year 1 for the stayers (see Table 20). Correlations were also conducted using alternate value-added scores (a) including only one year of prior achievement and (b) including class mean achievement. The correlations between these value-added scores and ratings of teaching practices were similar in magnitude.

Table 20. Correlations between Teachers' Value-added Scores and Practice Ratings for Stayers

	Fall Year 1 (n=28)	Spring Year 1 (n=28)	Spring Year 2 (n=28)	Spring Year 3 (n=28)
Overall IQA	0.08	0.09	0.29	0.52**
Emotional Support	0.09	0.00	0.33^{+}	0.08
Classroom Organization	0.22	0.20	0.29	0.04
Instructional Support	0.36^{+}	0.08	0.27	0.34^{+}

+ for p<.10, * for p<.05, ** for p<.01, and *** for p<.001

It is not surprising that classroom management practices may matter most for first year teachers because teachers' management skills are related to how much instructional content they are able to convey. Less time spent on classroom procedures and behavior management leads to higher productivity. Classroom environment practices not only "set the stage for all learning," but they also seem to "set the stage" for beginning teacher effectiveness (Danielson, 2007, p. 28). A possible explanation for declining correlations of value-added scores with classroom environment practices is that by year 3 most of the teachers that remained in teaching had become proficient at these practices, as indicated by the narrower standard deviations of Emotional Support and Classroom Organization in the third year. This occurrence would mean that high and low value-added teachers would have similar ratings of classroom environment practices, making the ratings less effective for identifying effective teachers.

The increasing correlations of value-added scores with instructional practices are more curious, especially given the finding that, on average, teachers did not improve their math-specific instructional practices. To dig deeper into this finding, I investigated whether change in ratings of teaching practice was correlated with change in value-added scores. We would expect a positive correlation if two trends were happening overall: (a) teachers who improved their practice also improved in their effectiveness at increasing student achievement and (b) teachers whose practice ratings diminished also had decreases in their value-added scores. To conduct this analysis, I created change scores for year 3 by subtracting year 2 scores from year 3 scores and change scores for year 2 by subtracting year 1 average scores from year 2 scores.

Improvements on IQA scores were positively associated with improvements in value-added scores in both years 2 and 3 (see Table 21). Therefore, while the growth curve analysis showed that on average teachers were not improving their IQA ratings, it appears that some teachers did improve their math-specific instructional practices and that their improvement was positively correlated with improvement in teacher effectiveness. In fact, of the 26 teachers with scores for year 2 and year 3, 9 teachers improved in both their Overall IQA rating and their value-added scores from year 2 to year 3, and 7 declined in both their Overall IQA rating and their value-added scores.

Table 21. Correlations between Change in Value-added Scores and Change in Practice Ratings

	Year 1 Average to Spring Year 2 (n=40)	Spring Year 2 to Spring Year 3 (n=26)
Overall IQA	0.21	0.38**
Emotional Support	0.19	-0.07
Classroom Organization	0.09	-0.12
Instructional Support	0.21	-0.06

+ for p<.10, * for p<.05, ** for p<.01, and *** for p<.001

In addition to examining the correlations at each time period, the data were also stacked across all time periods and correlations between value-added scores and practice ratings were conducted. This was done to compare the findings to those of other studies. No other studies have examined the relationship of value-added scores and ratings of teacher practice for teachers in each of their first few years of teaching, but pooling the data across the first three years allows me to examine the correlation between value-added scores and practices for beginning teachers more generally. Overall correlations were statistically significant and similar in size (see last row of Table 19). Correlations were the largest for the Overall IQA and smallest for Emotional Support. In general, the correlations between ratings of teaching practices and teacher value-added scores were fairly similar to those found in other studies (see Table A1 in Appendix). For example, Hill et al. (2010) found a correlation of 0.36 between math-specific practices measured by the MQI and teachers' value-added and Kane and Staiger (2012) found a 0.16 correlation for MQI ratings and value-added scores. The 0.26 correlation between Overall IQA ratings across time and value-added scores found in this study aligns with those findings.

To compare the correlations found between ratings of teacher practice and value-added scores found in this study to those found in the MET project, I calculated an overall CLASS-S score. Correlations between the Overall CLASS-S score across all time periods and teacher value-added scores from this study was 0.22, which is very similar to the 0.24 correlation found for math teachers in the MET study. This suggests that the relationship between teaching practices and teacher effectiveness does not vary greatly for teachers of different experience levels when using a combined measure of practices.

To further explore the relationship between ratings of teaching practice and a measure of teacher effectiveness, I calculated correlations with each of the dimensions of the CLASS-S and the individual IQA rubric scores for each time period (see Table 22). Given the results reported earlier, it is not surprising that in year 1 the CLASS-S dimension ratings with the highest correlations with value-added were behavior management, productivity, and instructional learning formats, which are part of the Classroom Organization domain. When compared to the Classroom Organization dimensions, the Instructional Support dimensions of quality of feedback and instructional dialogue are similarly correlated with value-added scores in fall of year 1.

However, in the spring of year 1 the value-added scores are not correlated with the Instructional Support dimension ratings. This indicates inconsistency of ratings from fall to spring of year 1 for the beginning teachers.

In years 2 and 3, a similar story appears across the two measures. Teachers who more frequently engaged students in higher-level thinking skills and problem-solving activities had higher contributions to student achievement. In years 2 and 3, Analysis and Problem Solving rankings were most highly correlated with value-added scores of all the CLASS-S dimension ratings (Year 2: r=0.41, p<.01; Year 3: r=0.37, p<.10). Of the three IQA Academic Rigor rubrics, Task Implementation was most highly correlated the highest with teacher value-added scores across all time periods (Year 2: r=0.18, p=0.24; Year 3: r=0.52, p<.01). As a reminder, Analysis and Problem Solving captures whether the teacher is providing opportunities for inquiry and analysis via open-ended and challenging tasks and asking students to explain their thinking. Like Analysis and Problem Solving, Task Implementation captures whether teachers engaged students with open-ended tasks, multiple representations of concepts, and making connections among ideas.

Table 22. Correlations between Value-added Scores and Ratings of Teachers' Practices at the Dimension and Rubric Level

	Fall Year 1 (n=57)	Spring Year 1 (n=56)	Spring Year 2 (n=42)	Spring Year 3 (n=28)
CLASS-S Dimensions				
Positive Climate	0.16	0.09	0.35^{*}	0.11
Negative Climate	0.10	0.02	0.35^{*}	-0.17
Teacher Sensitivity	0.08	0.08	0.30^{+}	0.09
Regard for Adolescent Perspective	0.20	0.09	0.05	0.17
Productivity	0.20	0.25^{+}	0.31^{*}	-0.11
Behavior Management	0.21	0.28^{*}	0.28^{+}	0.11
Instructional Learning Formats	0.28^{*}	0.18	0.22	0.13
Content Understanding	0.17	0.17	0.23	0.00
Analysis and Problem Solving	0.11	-0.03	0.41^{**}	0.37^{+}
Quality of Feedback	0.20	0.00	0.15	0.30
Instructional Dialogue	0.23	0.04	0.21	0.37
Student Engagement	0.26	0.20	0.29^{+}	0.23
IQA Rubrics				
Task Potential	0.05	0.12	0.12	0.39^{*}
Task Implementation	0.29^{*}	0.17	0.18	0.52^{**}
Discussion	-0.05	-0.16	0.10	0.41^{*}

+ for p<.10, * for p<.05, ** for p<.01, and *** for p<.001

Next, I investigated the magnitude of the differences in teachers' practice ratings between teachers who were most and least effective at increasing student achievement. Table 23 shows the teaching practice ratings for top quartile and bottom quartile value-added teachers in each time period. The difference teaching practice ratings between least effective and most effective teachers decreased over time for classroom environment practices and increased over time for instructional practices. In both time periods of year 1, the difference in Classroom Organization scores between least effective and most effective teachers was about a full point, whereas in year 3 the difference was about one-fifth of a point or less (see Table 23). In contrast, the difference in Instructional Support ratings was half a point or less in the first year and was two-thirds of a point in year 3. The difference in Overall IQA scores of top and bottom value-added teachers also increased over time.

Table 23. Teaching Practice Ratings by Value-added Score Quartiles

	Fall Year 1			Spring Year 1			Spring Year 2			Spring Year 3		
	Top Quartile (n=13)	Bottom Quartile (n=14)	Difference	Top Quartile (n=13)	Bottom Quartile (n=13)	Difference	Top Quartile (n=9)	Bottom Quartile (n=11)	Difference	Top Quartile (n=6)	Bottom Quartile (n=7)	Difference
Emotional Support	4.64	4.11	0.53*	4.51	4.03	0.48	4.91	4.50	0.41	5.09	4.94	0.15
Classroom Organization	4.94	4.03	0.91*	5.07	3.95	1.12**	5.34	4.78	0.56+	5.35	5.12	0.23
Instructional Support	3.80	3.32	0.48+	3.72	3.38	0.34	3.93	3.63	0.30	4.30	3.65	0.65
Overall IQA	1.71	1.78	-0.07	1.58	1.56	0.02	1.76	1.43	0.33	2.10	1.38	0.72

+ for p<.10, * for p<.05, ** for p<.01, and *** for p<.001

Finally, using hierarchical linear modeling with value-added scores and observation ratings nested within teacher, I found additional evidence of a significant relationship between Overall IQA ratings and value-added scores. Table 24 shows the estimated increase in value-added scores for a one unit increase in teachers' CLASS-S and Overall IQA scores. In models that included just individual practices (model 1), Classroom Organization (β = 0.017, p<0.05) and Overall IQA (β = 0.053, p<0.01) ratings were significantly associated with value-added scores, and Instructional Support ratings (β = 0.020, p<0.10) were marginally significantly associated with value-added ratings. I also tested whether the strength of the relationship between practice ratings and value-added scores differed by year by including interactions between practice ratings and the time variables (model 2). When including the interactions, it appears that in year 3 value-added scores decreased (β = -0.188, p<0.05) but that the decrease was significantly reduced if teachers improved their Overall IQA ratings (β = 0.121, p<0.05).

Table 24. Predicted Association between Value-Added Scores and Teaching Practice Ratings

	Emotional Support		Classroom Organization		Instructional Support		Overall IQA	
	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
Practice Rating	0.015	0.006	0.017^{*}	0.014	0.020^{+}	0.004	0.053^{**}	0.002
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.03)
Year 2	0.000	-0.131	-0.004	-0.077	-0.001	-0.113	0.004	-0.094
	(0.02)	(0.10)	(0.02)	(0.09)	(0.02)	(0.09)	(0.02)	(0.07)
Year 3	0.016	0.040	0.013	0.081	0.014	-0.125	0.017	-0.188^{*}
	(0.02)	(0.13)	(0.02)	(0.14)	(0.02)	(0.11)	(0.02)	(0.08)
Year 2 *		0.028		0.015		0.030		0.060
Practice Rating		(0.02)		(0.02)		(0.02)		(0.04)
Year 3 *		-0.004		-0.013		0.036		0.121^{*}
Practice Rating		(0.03)		(0.03)		(0.03)		(0.05)
Constant	-0.07	-0.030	-0.08^{*}	-0.066	-0.08	-0.016	-0.09^{**}	-0.006
	(0.05)	(0.06)	(0.04)	(0.05)	(0.04)	(0.06)	(0.03)	(0.05)
Grade/School Change	X	X	X	X	X	X	X	X
AIC	-249.75	-247.92	-251.74	-248.95	-250.92	-249.40	-255.2	-258.15
BIC	-230.06	-222.61	-232.05	-223.64	-231.23	-224.09	-235.61	-232.84
N	123	123	123	123	123	123	123	123

+ for p<.10, * for p<.05, ** for p<.01, and *** for p<.001

As we would expect a variety of teaching practices to influence teacher effectiveness at increasing student achievement, I tested whether the results held when controlling for ratings of other practices. I do not include the measures of cross-subject and math-specific instructional practices in the same models because of their conceptual overlap. Columns 1 and 2 of Table 25 present results from the analysis including the cross-subject Instructional Support ratings, and columns 3 and 4 included the math-specific Overall IQA ratings. As in the prior analysis, I ran the models with and without time interactions. Again, Classroom Organization ratings are significantly associated with value-added scores (β = 0.032, p<0.05), but only when not controlling for Overall IQA ratings. Instructional Support ratings were significantly associated with value-added scores significant relationship but only in year 3. Even when controlling for ratings of classroom environment practices, Overall IQA scores were significantly associated with value-added scores in year 3 (β = 0.140, p<0.01).

Table 25. Predicted Association between Value-Added Ratings and Teaching Practice Rating, Controlling for Other Practice Ratings

	(1)	(3)	(2)	(4)
Emotional Support	-0.014	-0.013	-0.008	-0.021
	(0.02)	(0.03)	(0.02)	(0.02)
Classroom Organization	0.018	0.032*	0.017	0.027
	(0.01)	(0.02)	(0.01)	(0.01)
Instructional Support	0.017	-0.019		
	(0.02)	(0.03)		
Overall IQA			0.048*	-0.005
			(0.02)	(0.03)
Year 2	-0.005	-0.156	-0.003	-0.196
	(0.02)	(0.10)	(0.02)	(0.10)
Year 3	0.012	0.077	0.010	-0.089
	(0.02)	(0.14)	(0.02)	(0.14)
Year 2 * Emotional Support		0.029		0.044
		(0.05)		(0.04)
Year 3 * Emotional Support		-0.029		0.017
		(0.05)		(0.04)
Year 2 * Classroom Organization		-0.018		-0.021
		(0.03)		(0.03)
Year 3 * Classroom Organization		-0.079		-0.043
		(0.04)		(0.04)
Year 2 * Instructional Support		0.027		
		(0.04)		
Year 3 * Instructional Support		0.122**		
		(0.04)		
Year 2 * Overall IQA				0.056
				(0.04)
Year 3 * Overall IQA				0.140**
				(0.05)
Grade/School Change	X	X	X	X
Constant	-0.081	-0.023	-0.121*	-0.026
	(0.05)	(0.06)	(0.05)	(0.06)
AIC	-248.586	-246.983	-253.532	-253.204
BIC	-223.277	-204.800	-228.223	-211.021
N	123	123	123	123

+ for p<.10, * for p<.05, ** for p<.01, and *** for p<.001

To further test the relationship between value-added and ratings of practice, I used a teacher fixed effects model. This method controls for time-invariant teacher characteristics. Only teachers with at least two years of value-added scores are included. The average teacher included in this analysis had 2.2 observations, meaning most teachers had data from two years. First, I included each of the practice ratings separately (Table 26). Then, I included classroom environment practices and instructional practices in the same models (Table 27). Only the relationship between the ratings of math-specific instructional practices was even marginally significantly associated with value-added scores. A one point gain in Overall IQA scores was associated with a 0.05 increase in value-added ($p<.05$). The standard deviation of value-added across time was 0.09 so this predicted increase represents about half of a standard deviation. As in the HLM models, it appears that the relationship between Overall IQA scores and value-added is largely driven by year 3 scores. The relationship between Overall IQA scores and value-added in year 3 remains marginally significant when controlling for classroom environment practice ratings ($\beta = 0.115$, $p<0.10$).

Table 26. Predicted Association between Value-Added Scores and Teaching Practice Ratings, Controlling for Time-Invariant Teacher Characteristics (Fixed Effects)

	Emotional Support		Classroom Organization		Instructional Support		Overall IQA	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Practice	-0.004	-0.017	-0.014	-0.027	0.002	-0.015	0.053^{+}	0.003
	(0.02)	(0.03)	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)	(0.04)
Year 2	0.015	-0.125	0.020	-0.124	0.013	-0.080	0.013	-0.055
	(0.02)	(0.12)	(0.02)	(0.10)	(0.02)	(0.11)	(0.02)	(0.09)
Year 3	0.036	0.049	0.040	0.111	0.034	-0.067	0.031	-0.141
	(0.02)	(0.16)	(0.02)	(0.16)	(0.02)	(0.13)	(0.02)	(0.10)
Year 2 * Practice Rating		0.030		0.030		0.025		0.042
		(0.03)		(0.02)		(0.03)		(0.05)
Year 3 * Practice Rating		-0.001		-0.012		0.026		0.101^{+}
		(0.03)		(0.03)		(0.03)		(0.06)
Grade/School Change	X	X	X	X	X	X	X	X
Constant	0.009	0.069	0.057	0.113	-0.015	0.047	-0.094*	-0.014
	(0.09)	(0.11)	(0.07)	(0.09)	(0.07)	(0.10)	(0.04)	(0.07)
Observations	123	123	123	123	123	123	123	123

+ for p<.10, * for p<.05, ** for p<.01, and *** for p<.001

Table 27. Predicted Association between Value-Added Scores and Teaching Practice Ratings, Controlling for Other Practice Ratings and Time-Invariant Teacher Characteristics (Fixed Effects)

	(1)	(3)	(2)	(4)
Emotional Support	0.005	0.020	0.010	0.022
	(0.04)	(0.05)	(0.03)	(0.04)
Classroom Organization	-0.028	-0.033	-0.020	-0.033
	(0.02)	(0.03)	(0.02)	(0.03)
Instructional Support	0.019	-0.012		
	(0.03)	(0.04)		
Overall IQA			0.051	0.002
			(0.03)	(0.05)
Year 2	0.020	-0.143	0.019	-0.154
	(0.02)	(0.13)	(0.02)	(0.13)
Year 3	0.040	0.159	0.036	0.027
	(0.02)	(0.18)	(0.02)	(0.18)
Year 2 * Emotional Support		0.031		0.020
		(0.06)		(0.05)
Year 3 * Emotional Support		-0.038		-0.018
		(0.06)		(0.05)
Year 2 * Classroom Organization		0.010		0.008
		(0.04)		(0.04)
Year 3 * Classroom Organization		-0.044		-0.017
		(0.05)		(0.05)
Year 2 * Instructional Support		-0.007		
		(0.05)		
Year 3 * Instructional Support		0.077		
		(0.06)		
Year 2 * Overall IQA				0.025
				(0.06)
Year 3 * Overall IQA				0.115^{+}
				(0.06)
Grade/School Change	X	X	X	X
Constant	0.026	0.095	-0.043	0.041
	(0.09)	(0.12)	(0.10)	(0.12)
N	123	123	123	123

+ for p<.10, * for p<.05, ** for p<.01, and *** for p<.001

When examining the association between change in value-added scores and change in teaching practice ratings, I found that a one point increase Overall IQA ratings is associated with an increase of 0.05 in value-added. The standard deviation for one-year change in value-added was 0.10 so a 0.05 change in value-added would be a change of half a standard deviation. This is more evidence for the relationship between improvements in Overall IQA ratings and increases in value-added scores and reflects the significant correlation found between value-added change scores and Overall IQA practice changes scores from year 2 to year 3 (Table 21). Practically speaking, teachers who get better at their math-specific instructional practices are significantly likely to increase their value-added scores.

Table 28. Predicted Association between Change in Value-Added Scores and Change in Teaching Practice Ratings

	Separate Models for Each Practice Rating				Multiple Practice Ratings Included	
	(1)	(2)	(3)	(4)	(5)	(6)
Emotional	0.002				0.026	0.035
Support	(0.02)				(0.03)	(0.03)
Classroom		-0.012			-0.035^{+}	-0.032^{+}
Organization		(0.01)			(0.02)	(0.02)
Instructional			0.003		0.013	
Support			(0.02)		(0.03)	
Overall IQA				0.050^{*}		0.049^{*}
				(0.02)		(0.02)
Grade/School						
Change	X	X	X	X	X	X
Constant	0.017	0.020	0.017	0.015	0.022	0.019
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
AIC	-105.434	-106.335	-105.448	-109.886	-104.334	-108.578
BIC	-94.485	-95.386	-94.500	-98.938	-89.007	-93.251
N	66	66	66	66	66	66

+ for p<.10, * for p<.05, ** for p<.01, and *** for p<.001

Using a variety of methods to examine the relationship between ratings of beginning middle school math teachers' classroom environment and instructional practices and their effectiveness and increasing student achievement, I found consistent evidence that teachers with higher ratings of their math-specific instructional practices and teachers who improved in their math-specific instructional practices saw greater increases in student learning. This relationship was largely driven by scores in teachers' third year. The following section discusses the implications of this finding and others for supporting beginning teachers and evaluating their practices.

CHAPTER V

DISCUSSION

This dissertation addresses gaps in existing teacher quality research by examining the development of beginning middle school math teachers' classroom environment and cross-subject and math-specific instructional practices, the relationship between these practices, and the association of these practices to teacher effectiveness. Findings support the notion that teaching is a complex practice. Teachers can be strong in some areas and weak in others, and some practices appear to leverage stronger gains in student learning. In this final chapter, I review the limitations and key findings of my study, discuss implications for policy and practice, and make suggestions for future research in this area.

Considerations

Before discussing the findings or implications, it is important to remember that the complex nature of teaching means there are a variety of ways that the quality of teaching practices can be measured. No standardized definition of what quality teaching looks like exists, which makes studying teaching practices challenging. Though this study captures a wide range of teaching practices, it is likely that there are still some aspects of teaching that impact student achievement that are not included. For example, neither the CLASS-S nor the IQA measures teachers' assessment practices which some studies have shown to be linked to student achievement (Wenglinksy, 2002).

Furthermore, with no standardized definition of teaching quality, different measures of teaching practices set their own bar for what is considered high quality. Different measures also set their own expectations for different levels of practice proficiency. In order to capture multiple dimensions of teaching practice, I used two different measures. Caution should be used when comparing practice ratings across measures as the conceptual nature of the measures and scales used varies.

Ratings from observational measures of teaching practice function as a proxy for actual quality of teaching. Likewise, student scores on state tests serve as a proxy for student achievement just as value-added estimates are a proxy for a teachers' effectiveness at increasing their students' achievement. As with any proxy measure, these ratings are likely to contain measurement error. There are several potential sources of measurement error in ratings of teaching practices. First, rater bias could occur if a rater is systematically assigning higher ratings to certain teachers. In this study, procedures like rater training and double-coding discussed in the methods chapter were put into place to reduce the likelihood of rater bias.

Next, for the purposes of this study, observations conducted during two days are assumed to be representative of a teachers' quality of teaching practice at a particular time period. Ideally, when assessing the trajectories of beginning teachers' instructional quality, observations would be conducted at different time points over multiple years, but measuring beginning teachers' instructional quality was one of many measures included in the AIM study and gaining multiple observation points per year was not the original focus of the study. Instruction may vary from day to day and across the school year and could be related to which standards and skills a teacher is covering. For example, as standardized achievement testing time approaches teachers may reduce the rigor of their instructional practices in favor of test preparation activities. If this is the case, observations in the spring may be systematically lower than observations in the fall or winter. Thus, relying solely on spring observations may mean that estimates of teaching practice growth are biased downward and might appear larger if observations took place over the course of the year. We did attempt to conduct the spring observations in February or early March or after testing to avoid this problem, but this was not always possible.

If all teachers similarly lowered their instructional rigor then this is not as problematic, but if teachers with high-level practices in the fall lowered their practices to the same level as lower-level teachers during the spring then relying solely on spring observations means we are not picking up on the differences in teaching practices from the fall. Moderate correlations in for teachers' ratings in the fall and spring of the first year[12] provide some evidence of a consistency of ratings for individual teachers across the school year.

A final limitation is that we are unable to account for extra instruction, like tutoring, that may be occurring. If teachers with lower ratings of teaching practices compensated by offering greater amounts of tutoring either before or after school or during free periods, estimates of the relationship between ratings of teaching practices and student achievement could be biased downward because teachers with lower practice ratings had higher student achievement gains driven by instruction occurring outside of the classroom. On the other hand, if teachers who have higher ratings of their teaching practices are also the teachers more likely to provide their students with tutoring, estimates could be biased upward. The same logic applies for other forms of math instruction that could be occurring outside of the math teacher's classroom such as math instruction during a science class or even parents providing math instruction at home. If this extra instruction was occurring more frequently for students of teachers with either low or high quality teaching practices, estimates could be biased.

Review of Findings

Now that I have noted a few limitations to consider, I review the results and provide possible explanations for my findings.

[12] Correlations for teachers' ratings in the fall and spring of their first year ranged from 0.42 to 0.59 for the CLASS-S and 0.17 for Overall IQA ratings.

Development of Beginning Teachers' Practices

Beginning middle school math teachers in this study started their careers with higher ratings of classroom environment practices than instructional practices. Studies examining teaching practices across experience levels have also found that ratings of teaching practices are typically higher for indicators of classroom environment practices than for instructional practices (Grossman et al., 2010; Kane & Staiger, 2012). Likewise, the "mid-level" ratings of beginning teachers' teaching practices found in this study are similar to the ratings reported by other studies using the CLASS-S and the IQA to assess teaching practices (Bell et al., 2012; Grossman et al., 2010; Matsumura et al., 2006).[13]

The beginning teachers in this study were predicted to improve most on their classroom management practices – by about a third of a point in a year and almost a point over three years. Moving from a 5 to a 6, for example, would indicate that teacher goes from occasional use of effective methods to encourage desirable behavior and prevent misbehavior to more consistently demonstrating clear expectations and proactive monitoring and that there are fewer instances of student misbehavior. The beginning math teachers also had significant growth on their social/emotional practices and their cross-subject instructional practices. They were predicted to improve about two-thirds of a point on their Emotional Support and Instructional Support ratings. For Emotional Support, moving from a 5 to a 5.66 would mean a teacher goes from demonstrating some positive communications, responsiveness to academic and social/emotional

[13] Bell et al. (2012) reported average CLASS-S ratings for about 400 Algebra I lessons taught by teachers across experience levels. They reported average ratings of 5.67 for Classroom Organization, 4.00 for Emotional Support, and 3.61 for Instructional Support. Grossman et al. (2010) reported averages in the "mid" range for the Emotional Support and Classroom Organization dimensions of positive climate (4.42), regard for adolescent perspective (3.45), behavior management (4.59), and productivity (4.31). Matsumura et al. (2006) reported "mid-level" average ratings for Task Potential (2.46) and Task Implementation (2.28), and lower average ratings for Discussion (1.65).

needs, and opportunities for student autonomy and leadership to engaging in these behaviors almost most of the time. Moving another two-thirds of a point from a 5.66 to a 6.33 would mean the teacher is now demonstrating these behaviors most nearly all of the time. For Instructional Support, moving from a 3.33 to 4 would mean a teacher would move from infrequent to occasional opportunities for higher level thinking and content-driven conversations.

This study does not address why the beginning teachers improved their management practices more than their instructional practices, but existing literature studying beginning teachers provides possible hypotheses. First, beginning teachers tend to be more concerned about their classroom management than their instruction (e.g., Oberski et al., 1999). Because of this, they likely focus more time on working to improve their classroom routines and procedures and their ability to encourage on-task student behaviors. Next, beginning teachers tend to receive more support around their classroom environment practices than their instructional practices (Hobson et al., 2009; Pourdavood et al., 1999). Principals and mentors often focus on mastery of management practices as a key indicator of success for beginning teachers. This is likely because good classroom management is viewed as a precursor to delivering quality instruction. Thirdly, improving instructional practices may just be more challenging, especially since beginning teachers have been found to lack adequate content knowledge and pedagogical content knowledge (Grossman, 1992; Kauffman, Johnson, Kardos, Liu, & Peske, 2002).

While the beginning math teachers experienced statistically significant improvement in the areas of classroom environment and cross-subject instructional practices, they were not predicted to improve their math-specific instructional practices. Across time, most teachers assigned and implemented procedural math lessons with discussion characterized by one-word student responses. There are few potential explanations for why beginning teachers showed improvement on cross-subject instructional practices, but not their math-specific instructional practices.

First, the measures used to assess these practices capture different aspects of instructional practices. The IQA rates teachers on the degree to which they assign and implement cognitively demanding problem-solving math tasks and organize discussion emphasizing reasoning and connections among mathematical ideas, whereas the Instructional Support assesses the extent to which teachers use a variety of delivery, assignment, and discussion approaches to help students understand whatever content is being taught. A teacher could improve in some of these practices, while not improving the others. For example, the average Instructional Support score for teachers implementing non-rigorous math instruction improved from a 3.4 to a 3.9 over time (see Figure 5). This shows that even though teachers were implementing procedural math lessons, they demonstrated more frequent use of the practices assessed by the Instructional Support domain, such as their communication of concepts and procedures, transmission of content knowledge, or attention to background knowledge. Furthermore, teachers' scores on Instructional Support can be boosted by giving students the opportunity to practice procedures and skills (Content Understanding dimension) and by content-based exchanges occurring within the lesson set-up or student work time (Quality of Feedback and Instructional Dialogue dimensions), whereas as the IQA Discussion rubric focuses exclusively on discussion occurring after student work time.

Relationship between Teaching Practices

Despite differences in improvement on classroom environment and cross-subject and math-specific instructional practices, in general, teachers who rated highly on one area of practice also had higher scores on other areas of practice. Beginning teachers who demonstrated higher levels of mastery of their classroom environment practices (i.e., Level 5 or higher on

Emotional Support and Classroom Organization) were significantly more likely to both assign and implement rigorous math instruction. Similarly, another study using the IQA found that the presence of explicit rules in the classroom for respectful, prosocial behavior significantly predicted the number of students who participated in discussions (Matsumura, Slater, & Crosson, 2008).

Concerns about classroom management may constrain beginning teachers from implementing inquiry-based instruction (Roerhing & Luft, 2004). Curricula like the Connected Math Project, which was used by 42% of the teachers in this study, includes group work activities where students are expected to be on-task and engaged in rigorous math activities without heavy teacher oversight. Teachers who felt their management was lacking may have been hesitant to assign these types of activities.

Association between Teaching Practices and Teacher Effectiveness

This study adds to the existing evidence that classroom observations can "capture elements of teaching that are related to student achievement" (Kane, 2010, p. 27). When analyzing the relationship between ratings of the teachers' practices and their value-added scores, it appears that instructional practices, especially those related to implementing rigorous math activities and discussion, are key. Findings indicate that classroom environment practices may be more important for increasing student learning in the first year of teaching before teachers become proficient at their instructional practices, but instructional practices are most correlated with teacher effectiveness ratings by the third year.

Teachers who more frequently engaged students in higher-level thinking skills and problem-solving activities had higher contributions to student achievement. This finding corresponds with other research that has advocated the need for more rigorous instruction (Stein

& Lane, 1996; NCTM, 2000). For example, an in-depth case analysis conducted by the National Center on Scaling Up Effective Schools found that a key difference between high and low value-added schools was the rigor of instructional practice, along with the rigor of student and teacher expectations (Smith, Taylor-Hayes, Preston, Vineyard, Katterfeld, & Neergaard; 2012). Additionally, teachers who improved in their math-specific instructional practices saw greater increases in student learning, with the relationship largely driven by scores in teachers' third year. These findings reflect the MET Project evidence for a stronger relationship between student achievement gains and ratings from subject-specific measures of teaching practices compared to ratings from cross-subject tools (Kane & Staiger, 2012).

Policy Implications

In this section, I discuss potential implications for improving the teacher workforce. I focus on policy levers that could be used to incentive actions that would result in the improvement of teachers' practices and, in turn, their students' achievement.

Teacher Evaluation

Recent federal legislation has pressured states and districts to improve teaching quality through teacher evaluation (United States Department of Education, 2009; United States Department of Education Office of Planning Evaluation and Policy Development, 2010). Evaluation systems are generally used for two purposes: (a) sorting teachers for high-stakes personnel decisions and (b) providing teachers feedback about the strengths and weaknesses of their teaching so that they can improve their teaching practices (Bell et al., 2012). Each of these functions is expected to increase student achievement.

Evaluation systems have traditionally included observational assessments completed by school and district officials. Recently, however, researchers and policymakers have been focused on other measures of teacher evaluation, such as value-added methods and student surveys (Klein, 2012). A great deal of attention is being given to what these measures should look like and how they should be included in teacher evaluation systems (Kane & Staiger, 2012; National Council on Teacher Quality, 2014). Still, observation measures remain the primary tool being used for evaluating teachers and providing them with feedback about their practices, with 28 states requiring annual observation evaluations of all teachers (National Council on Teacher Quality, 2014; Papay, 2012).

Research has found that providing teachers with feedback on their practice is associated with substantial improvements in teachers' practices (Chung, 2008; Wei & Pecheone, 2012) and student achievement gains in teachers' classrooms even without a targeted professional development effort (Allen et al., 2011; Taylor & Tyler, 2011). If an ultimate goal of evaluations is providing teachers with feedback so they can, then we need to think about how to better structure the content of evaluation tools to foster this process. After all, the practice constructs measured by the observation protocols will improve their teaching shape the feedback teachers receive. Kane & Staiger (2012) explained that when designing observational measures, we should be mindful of which competencies of teaching practice teachers are more likely to improve given appropriate feedback and training opportunities. They write, "we are better off measuring competencies that teachers are inspired to improve and can improve with the right supports" (p. 33).

Identifying areas of strength and areas for improvement is especially critical for beginning teachers who often feel as though they are failing (Le Maistre & Paré, 2010) and are more likely to stay in their schools if they feel successful (Johnson & Birkeland, 2003). Findings from this study suggest including classroom environment practices when evaluating beginning teachers is useful because new teachers are likely to show improvement that can help them feel successful. Evaluating instructional practices is also important to help identify areas that teachers can target for improvement, though it is important to keep from overwhelming new teachers with numerous areas of focus.

The practices evaluated by observation measures used in teacher evaluation not only shape the feedback that teachers receive, but they also shape the ratings teachers receive. Ratings are dependent on the aspects of teaching practice assessed. I found that teachers may rate highly in some areas and low in others. Therefore, if an evaluation measured only assessed instructional practices, teachers with good environment practices and weaker instructional practices would be disadvantaged.

Furthermore, only assessing a teacher's cross-subject instructional practices may present a limited view of their teaching practices. As Hill & Grossman (2013) write, "most of the observation protocols selected in new teacher evaluation systems are generic with respect to content area and are designed to be used with all teachers—from kindergarten through calculus" (p. 373). If effective teaching is considered to look differently across subject areas (Graeber, Newton, & Chambliss, 2012; Grossman & Stodolsky, 1995), then we should consider adding components that address teachers' subject-specific practices. This is especially relevant given the finding from this study that math-specific practices are more highly correlated with teachers' value-added scores. Overall, findings from this study remind us of the importance of carefully considering the implications of the practices assessed by observation measures, especially when the ratings are used to make high-stakes personnel decisions like job and licensure renewal.

Linking teacher evaluation to high-stakes decisions can incentivize certain behaviors. Given the finding that instructional practices related to promoting higher-order thinking are more highly correlated with student achievement gains, policymakers should consider assigning more weight to those practices when evaluating teachers. For example, in Tennessee, 50% of a teacher's overall evaluation score comes from observational evaluation ratings. The rubric most commonly used across the state includes 23 indicators: 3 for Planning, 4 for Professional Growth, 4 for Environment, and 12 for Instruction. Teachers are observed multiple times on a subset of indicators and then their scores are averaged to provide a single score. The larger number of Instruction indicators prioritizes these practices.

Training and Support

Teacher evaluation can help teachers improve by providing them systematic feedback about the quality of their teaching practices, but beginning teachers often need additional supports to help them improve their teaching. Beginning math teachers in this study had lower ratings on instructional practices than classroom environment practices. They also struggled to improve their math-specific instructional practices. Teacher educators and school systems should consider what can be done to help beginning teachers improve their instructional practices. This is especially important given the strong association found between ratings of instructional practices and student achievement gains. We need to better train all teachers on how to provide students with opportunities to complete open-ended tasks that engage them in higher-order thinking.

Content-focused pre-service training and beginning teacher supports such as mentoring or professional development may help with this goal. Darling-Hammond and McLaughlin (1995) attribute weaknesses in beginning teachers' pedagogical content knowledge to pre-service training that does not emphasize teachers' deep understanding of subject knowledge and how students learn that subject knowledge. Studies have found that subject-specific professional development and coaching are more effective in improving teachers' instructional practices (Biancarosa, Bryk, & Dexter, 2010; Cohen & Hill, 2001; Desimone, Porter, Garet, Yoon, & Birman, 2002). Roehrig and Luft (2006) found that beginning teachers receiving content-focused professional development experiences were more likely to implement reform-based lessons and student-centered practices. New teacher mentoring can also contribute to improvement in instructional practices, such as leading discussions (Stanulis, Little, & Wibbens; 2012). States and districts should reorganize supports to target the instructional needs of beginning teachers.

Standards and Curriculum

Grade-level content standards and curriculum are also policy levers states and districts can use to shape teaching practices. Standards are a set of content-specific learning goals for what a student should know and be able to do at the end of each grade. Recently, 45 states adopted the Common Core State Standards (CCSS). When compared with previous state standards for mathematics and English language arts and literacy, the CCSS have been found to place a greater emphasis on higher order cognitive demand (Porter, McMaken, Hwang, & Yang, 2011). Furthermore, Schmidt and Houang (2012) found a high degree of similarity between the Common Core Standard Standards in Mathematics (CCSSM) and the standards of the highest-achieving nations on an international math assessment and that states with prior standards more aligned to the CCSSM had higher performance on a national assessment of student achievement. Currently, there is some pushback on the implementation of the CCSS and CCSS-aligned assessments. Findings from this study suggest that staying the course with implementation of the CCSS would increase student achievement because teachers would be encouraged to deliver instruction that promotes higher-order thinking.

Inexperienced teachers are especially prone to adhere closely to the curriculum and materials that they are provided (Grossman & Thompson, 2008). Thus, making sure beginning teachers have access to high-level curriculum is especially important. Results from this study indicate that teachers who assign more challenging tasks have greater increases in their student achievement. Another study using AIM data found that having access to a reform mathematics curriculum was significantly associated with implementing more rigorous tasks (Smith, Neergaard, Hochberg, & Desimone, 2011). Moreover, findings from a number of studies conclude that students who have access to reform-oriented curricula perform better on conceptual understanding and problem solving than students taught with traditional curricula (Schoenfeld, 2002). States and districts should work to provide teachers with curriculum and instructional materials that will facilitate their implementation of rigorous instruction.

Future Research

I conclude with suggestions for future research motivated by the findings from this study. First, a number of questions arise from this study regarding the development of beginning middle school math teachers' practices. This study leaves us questioning (a) why the beginning teachers were more proficient at their classroom environment practices, (b) why they improved most on their classroom management practices, and (c) why they improved their cross-subject instructional practice but not their math-specific instructional practices. While I suggest some possible hypotheses to these questions earlier in the discussion section, more in-depth analysis is needed to address these questions. Deeper analysis of teacher observations, along with teacher interviews that ask about teachers' pre-training experiences, school supports, and notions of quality teaching, could provide insight to address these questions.

Next, while existing research has identified some supports that are associated with improvement in teaching practices and gains in student achievement, the field would benefit from more awareness of the supports that can help beginning teachers improve their cross-subject and math-specific instructional practices. Additionally, greater knowledge is needed about what supports can help all math teachers implement inquiry-based math instruction that promotes higher order thinking. As the AIM study collected information on the supports provided to the beginning teachers, we plan to contribute to the field by taking this study another step to examine what supports are associated with increases in ratings of different areas of teaching practices.

Finally, this paper draws on data from a longitudinal study of the mentoring and induction experiences of middle school mathematics teachers to investigate the development the beginning teachers' practices and the relationship of classroom environment and content-neutral and math instructional practices. While this was not an initial goal of the study, the recorded classroom observations were used to address a variety of research questions, such as alignment between instruction and supports and differences in instructional practices between stayers, leavers, and movers, as well as to address the research questions in this study. Researchers should be encouraged and allowed flexibility in making use of data to address a range of research questions. For example, evaluations of professional development programs and other educational interventions often gather longitudinal data on teacher practices and student achievement (e.g., (Desimone et al., 2002). This observation data has primarily been used to measure the effectiveness of interventions and educational programs, but it could be employed to study change in practices over a school year and the relationship of practices to student achievement. As teaching practices are the means through which student learning occurs, more research is needed that explores the development of teaching practices across experience levels and subject areas. We should take advantage of existing data to foster a better understanding of how teaching practices develop and change over time and which areas of practice are most associated with increases in student achievement

APPENDIX

Table A1. Students Examining the Relationship between Teaching Practices and Student Achievement

Combined Measures of Teaching Practices				
Authors	**Measure of Teaching Practices**	**Data**	**Methods**	**Main Findings**
Schacter & Thum (2004)	Combined measure of 12 practices: questions, feedback, presentation, lesson structure and pacing, lesson objectives, classroom environment, grouping students, thinking, activities, motivating students, and teacher knowledge of students	52 experienced teachers from 5 elementary schools were observed 8 times during 2001-02 by trained researchers	Correlation between a combined measure of teaching practices and teacher value-added scores	Math: r=0.55 Reading: r=0.68 Language arts: r=0.70
Holtzapple (2003)	Teacher evaluation scores were based on Danielson's (1996) *Framework for Teaching*, with specific focus on the Teaching for Learning domain	About 80 Cincinnati Public School teachers in grades 3-8 in 2000-01 and about 166 teachers in 2001-02 were observed 6 times during each year by district evaluators and building administrators	Correlations between composite evaluation ratings and mean classroom gains	Math: r=0.38 Reading: r=0.27 Science r=0.27 Social Studies: r=0.29
Borman & Kimball (2005)	Teacher's evaluation scores were based on Danielson's (1996) *Framework for Teaching*, with specific focus on the Planning and Preparing and Teaching for Learning domains.	About 400 teachers in grades 4-6 from Washoe County (Nevada) were observed. Beginning teachers were observed at least 9 times during a year. Experienced teachers are observed 1-3 times per year	Classroom level using hierarchical linear modeling to predict classroom mean achievement based on teacher's evaluation scores	A teacher with an evaluation score one standard deviation above the mean was associated with average classroom achievement scores one fifth of a standard deviation above scores of students taught by a teacher at one standard deviation below the mean
Milanowski (2004)	Teacher's evaluation scores were based on Danielson's (1996) *Framework for Teaching*, using all four domains.	212 Cincinnati Public School teachers in grades 3-8 in 2000-01 and 2001-02 were observed 6 times during each year by district evaluators and building administrators	Correlations between teacher evaluation scores and value-added scores	Math: r=0.43 Science: r=0.27 Reading: r=0.32
Heneman, Milanowski	Teacher's evaluation scores were based on Danielson's (1996)	Teachers in 4 districts (2,500 in Cincinnati; 40 in a Los Angeles	Correlations between teacher	Correlations varied across their four sites

, Kimball, and Odden (2006)	*Framework for Teaching*, using all four domains	charter school; 3,300 in Reno/Sparks, Nevada; and 475 in Coventry, Rhode Island) were observed by building administrators and district evaluators from 1999 to 2004	evaluation scores and value-added scores	Math: r=0.11 to 0.32 Reading: r=.22 to 0.37
Kane et al. (2010)	Teacher's evaluation scores were based on Danielson's (1996) *Framework for Teaching* and the evaluation scores are broken into three components based on factor analysis: classroom environment, teaching through questioning and discussion, and routinized standards and content focused teaching	2,071 Cincinnati Public School teachers in grades 3-8 in years 2001 to 2009 were observed between 1 to 8 times during each year by trained professionals external to the school	Separated teachers into quartiles based on their valued added scores and then conducted mean difference tests (t-tests) between evaluation scores of teachers in the different quartiles	A one point increase in average TES score is associated with a student achievement gain of about one-sixth of a standard deviation in math and one-fifth in reading. A one point increase in the average scores across the eight standards represents an increase of about two standard deviations. A teacher who scores higher on "classroom environment" (Domain 2) relative to "classroom practices" (Domain 3) is predicted to produce additional student gains; with coefficients of 0.25 standard deviations in math and 0.15 in reading.
Bell et al. (2012)	Teaching practices were measured by the CLASS-S domains of Emotional Support, Classroom Organization, and Instructional Support	82 Algebra teachers in 20 middle schools and 20 high schools in a district were observed 4 or 5 times during the same class period	Correlations between teacher evaluation scores and value-added scores	Emotional Support: 0.20 Classroom Organization: r = 0.25 Instructional Support: r = 0.19
Allen (2011)	Teaching practices were measured by the CLASS-S domains	78 secondary teachers in Virginia were randomly assigned to receive My Teaching Partner, a web-mediated professional development approach focused on improving teacher-student interactions	They used multilevel structural equation modeling to examine whether the student achievement gains from attending training were mediated by changes to teaching practices	Gains in student achievement on the state test in the year following the completion of the training appeared to be mediated by changes in teaching practices
Subject-specific Measures of Practices				
Grossman (2010)	Teaching practices were measured using PLATO, a	Grade 6, 7, and 8 teachers in their third through fifth years of	Separated teachers into quartiles of their	Teachers in the top quartile of value-added scores scored higher

	English Language Arts tool and the Emotional Support and Classroom Organization domains of CLASS-S	teaching in New York City were observed on 6 days during the spring of 2007 by outside researchers	valued added scores and then conducted mean difference tests (t-tests) between evaluation scores of teachers in the different quartiles	than second-quartile teachers on all 16 elements of instruction that were measured
Matsumura (2006)	Practices were measured using the math-specific Instructional Quality Assessment and the reading-specific Instructional Quality Assessment. Both assess the rigor of lesson activities and the quality of classroom discussion	21 ELA and 13 math sixth and seventh grade teachers from five urban middle schools were observed on 2 consecutive days for the same class period by the same outside rater over a 2 week period	Linear regression controlling for students' prior achievement and background characteristics to explore the relationship between the IQA ratings and achievement scores	Reading-specific IQA scores were significantly associated with students' achievement on the Reading Comprehension subscale of the SAT-10. Math-specific IQA scores significantly predicted student achievement on the Total Math subscale and the Procedures subscale
Hill, Kapitula, & Umland (2010)	Practices measured using the MQI, and math-specific tool	24 middle school math teachers had lessons observed on 6 days between January and March 2008 and rated by outside researchers	Correlations between teacher evaluation scores and value-added scores	MQI and value-added: r=0.36
MET Kane & Staiger (2012)	FFT (only Classroom Environment and Instruction domains), CLASS, ELA-specific PLATO, and math-specific MQI and UTeach Teacher Observation Protocol (UTOP)	Over 1,000 fourth through eighth grade teachers in 6 districts were video-taped and rated by outside evaluators	Correlations between teacher evaluation scores and value-added scores Separated teachers into valued added score quartiles and conducted mean difference tests (t-tests) between evaluation scores	Math CLASS: r=0.24 FFT: r=0.19 UTOP: r=0.26 MQI: r=0.16 ELA CLASS: r=0.10 FFT: r=0.11 PLATO: r=0.24

Table A2. Complete Results of Growth Models Predicting Change over Time in Beginning Teachers' Practices

	Overall IQA	Emotional Support	Classroom Organization	Instructional Support
Month	0.002	0.019^{***}	0.025^{***}	0.019^{***}
	(0.00)	(0.00)	(0.01)	(0.00)
Advanced Math	0.120	0.103	0.122	0.225
	(0.07)	(0.12)	(0.15)	(0.12)
Grade or School Change	0.134	0.029	0.005	-0.004
	(0.12)	(0.17)	(0.20)	(0.18)
Constant	1.583^{***}	4.333^{***}	4.498^{***}	3.349^{***}
	(0.06)	(0.11)	(0.16)	(0.11)
sd (month)	0.000	0.001^{*}	0.013^{***}	0.002
	(0.00)	(0.00)	(0.01)	(0.00)
sd(_cons)	0.208^{***}	0.556^{***}	0.897	0.510
	(0.06)	(0.09)	(0.12)	(0.00)
cor(month, _cons)	1.000	-1.000	-1.000	-1.000
	(0.00)	(0.00)	(0.00)	(0.00)
sd(residual)	0.369^{***}	0.524^{***}	0.656^{***}	0.564
	(0.02)	(0.03)	(0.04)	(0.00)
AIC	225.956	409.433	506.672	415.483
BIC	252.222	435.698	532.938	428.615
N	197	197	197	197

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A3. Complete Results of Growth Models Predicting Change over Time in Beginning Teachers' Practices for Teachers Remaining in Study All Three Years

	Overall IQA	Emotional Support	Classroom Organization	Instructional Support
Month	0.004	0.018***	0.024***	0.017***
	(0.00)	(0.00)	(0.01)	(0.00)
Advanced Math	0.164*	0.004	0.027	0.127
	(0.08)	(0.12)	(0.15)	(0.13)
Grade or School Change	0.172	-0.090	-0.080	-0.077
	(0.12)	(0.18)	(0.21)	(0.19)
Constant	1.536***	4.419***	4.521***	3.517***
	(0.08)	(0.12)	(0.17)	(0.13)
sd (month)	0.000	0.002***	0.007***	0.003***
	(0.00)	(0.00)	(0.00)	(0.00)
sd(_cons)	0.222	0.415***	0.664**	0.398***
	(0.00)	(0.08)	(0.09)	(0.11)
cor(month, _cons)	1.000	1.000	-1.000	1.000
	.	(0.00)	(0.00)	(0.01)
sd(residual)	0.348	0.514***	0.632***	0.537***
	(0.00)	(0.04)	(0.05)	(0.04)
AIC	141.309	271.108	326.914	280.072
BIC	152.989	294.468	350.274	303.432
N	137	137	137	137

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A4. Predicted Probability of "High" Instructional Practices for Prior Classroom Environment Practices from the Same Observation Period (HLM)

	High Task Potential (3 or 4)		High Task Implementation (3 or 4)	
	Classroom Organization	Emotional Support	Classroom Organization	Emotional Support
Month	0.008	0.006	0.005	-0.000
	(0.02)	(0.01)	(0.02)	(0.02)
Level 4 Average	0.549	0.618	0.712	0.801
	(0.51)	(0.52)	(0.64)	(0.72)
Level 5 Average	1.248*	1.591**	1.210*	1.907**
	(0.49)	(0.52)	(0.61)	(0.71)
Level 6 or 7 Average	1.433**	1.944**	1.581**	2.641***
	(0.49)	(0.60)	(0.61)	(0.76)
Advanced Math	0.710*	0.660*	0.688*	0.610
	(0.31)	(0.31)	(0.34)	(0.34)
% School FRPL	-0.022	-0.026	-0.001	-0.003
	(0.03)	(0.03)	(0.02)	(0.02)
Reform Math Curriculum	1.149***	1.082**	0.861*	0.769*
	(0.33)	(0.33)	(0.36)	(0.36)
Constant	-2.305***	-2.407***	-3.045***	-3.350***
	(0.53)	(0.56)	(0.65)	(0.75)
sd (month)	0.053***	0.048***	0.046***	0.043***
	(0.02)	(0.02)	(0.03)	(0.03)
sd(_cons)	0.786	0.664	0.871	0.796
	(0.39)	(0.43)	(0.48)	(0.50)
cor(month,_cons)	-0.526	-0.351	-0.485	-0.422
	(0.39)	(0.61)	(0.51)	(0.64)
AIC	501.613	494.013	435.766	421.594
BIC	545.156	537.555	479.308	465.137
N	387	387	387	387

Table A5. Predicted Probability of “High” Instructional Practices for Prior Classroom Environment Practices from the Prior Observation Period (HLM)

	High Task Potential (3 or 4)		High Task Implementation (3 or 4)		High Instructional Support (5, 6, or 7)	
	Classroom Organization	Emotional Support	Classroom Organization	Emotional Support	Classroom Organization	Emotional Support
Month	0.009	-0.000	0.037	0.023	0.059	0.070
	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)	(0.04)
Level 6 or 7 Average	-0.172	1.392*	-0.522	1.146	-0.055	-1.045
	(0.42)	(0.61)	(0.50)	(0.62)	(0.57)	(0.94)
Advanced Math	0.702	0.603	1.010*	0.884	0.563	0.651
	(0.45)	(0.43)	(0.51)	(0.49)	(0.56)	(0.60)
% School FRPL	-0.014	-0.021	0.055	0.042	0.055	0.058
	(0.05)	(0.05)	(0.07)	(0.06)	(0.07)	(0.08)
Reform Math Curriculum	1.058*	0.958*	0.845	0.659	1.254	1.407
	(0.45)	(0.42)	(0.53)	(0.48)	(0.66)	(0.74)
Constant	-1.177*	-1.112	-2.721***	-2.575**	-4.085***	-4.466***
	(0.59)	(0.57)	(0.82)	(0.79)	(0.94)	(1.08)
sd (month)	0.105***	0.113***	0.100***	0.110***	0.076***	0.081***
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
sd(_cons)	2.301**	2.333**	2.879***	2.840***	2.336	2.633*
	(0.74)	(0.73)	(0.86)	(0.85)	(1.02)	(1.10)
cor(month,_cons)	-0.929***	-0.951***	-0.943**	-0.963***	-0.804	-0.783
	(0.06)	(0.04)	(0.06)	(0.04)	(0.21)	(0.22)
AIC	360.492	355.148	306.965	304.731	245.842	244.414
BIC	392.709	387.366	339.183	336.949	277.923	276.495
N	265	265	265	265	261	261

Table A6. Value-added Scores

	Mean	Standard Deviation	Minimum	Maximum	Quartile 1	Quartile 2	Quartile 3	Quartile 4
Year 1	-0.004	0.07	-0.22	0.21	-0.08	-0.01	-0.00	0.08
Year 2	0.000	0.10	-0.20	0.44	-0.10	0.02	0.01	0.12
Year 3	0.007	0.10	-0.21	0.34	-0.10	-0.01	0.02	0.12
Across Time	0.000	0.09	-0.22	0.44				

Table A7. Value-added Scores for Stayers

	Mean	Standard Deviation	Minimum	Maximum
Year 1	-0.011	0.08	-0.22	0.14
Year 2	-0.014	0.08	-0.20	0.22
Year 3	0.007	0.10	-0.21	0.34

REFERENCES

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools, *Journal of Labor Economics*, *25*(1), 95-135.

Allen, J.P., Pianta, R.C., Gregory, A., Mikami, A.Y., & Lun, J. (2011). An interaction-based approach to enhancing secondary school instruction and student achievement. *Science* 333(6045):1034-37.

Amrein-Beardsley, A. (2008). Methodological concerns about the education value-added assessment system. *Educational Researcher, 37*(2), 65-75.

Ball, D. L. (1990). The mathematical understandings that prospective teachers bring to teacher education. *The Elementary School Journal*, 449-466.

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

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

Berliner, D. C. (1988). Implications of studies on expertise in pedagogy for teacher education and evaluation. *New directions for teacher assessment*, 39-68.

Biancarosa, G., Bryk, A. S., & Dexter, E. R. (2010). Assessing the value-added effects of literacy collaborative professional development on student learning. *The Elementary School Journal*, *111*(1), 7-34.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (2000). *How people learn*. Washington, DC: National Academy Press.

Britt, A. (1997). *Perceptions of beginning teachers: Novice teachers reflect upon their beginning experiences.* (ERIC Document Reproduction Service No. ED415218).

Boe, E. E., Shin, S., & Cook, L. H. (2007). Does teacher preparation matter for beginning teachers in either special or general education? *The Journal of Special Education*, *41*(3), 158-170.

Borko, H., & Putnam, R.T. (1996). Learning to teach. *In* D. Berliner *&* R. Calfee *(Eds.), Handbook of educational psychology (pp.* 673-708*).* New York, NY*:* Macmillan*.*

Borman, G. D., & Kimball, S. M. (2005). Teacher quality and educational equality: Do teachers with higher standards-based evaluation ratings close student achievement gaps? *The Elementary School Journal*, *106*(1), 3-20.

Boston, M. & Wolf, M. K. (2006). Assessing academic rigor in mathematics instruction: The development of Instructional Quality Assessment Toolkit. National Center for Research on Evaluation, Standards, and Student Testing (CRESST) Report #672.

Brophy, J. E., & Good, T. (1986). Teacher behavior and student achievement. In M. Wittrock (Ed.), *Handbook of research on teaching* (pp. 328-375). New York: Macmillan.

Brummet, Q., & Gershenson, S. (2012). Self-contained classroom teacher grade-level reassignments: Evidence from Michigan. Presented at Association for Education Finance and Policy.

Cheney, C., Krajewski, J., & Combs, M. (1992). Understanding the first year teacher: Implications for induction programs. *Teacher Education and Special Education, 15*(1), 18-24.

Chung, R. R. (2008). Beyond assessment: performance assessments in teacher education, *Teacher Education Quarterly, 35* (1): 7-28.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2007). Teacher credentials and student achievement: Longitudinal analysis with student fixed effects. *Economics of Education Review*, *26*(6), 673-682.

Cohen, D. K., McLaughlin, M., & Talbert, J. (1993). *Teaching for understanding: Challenges for practice, research and policy*. San Francisco: Jossey Bass.

Cohen, D. K., & Ball, D. L. (1990). Relations between policy and practice: A commentary. *Educational Evaluation and Policy Analysis, 12*(3), 331-338.

Cohen, D.K., & Hill, H.C. (2001). *Learning policy: When state education reform works*. New Haven, CT: Yale University Press.

Coker, H., Medley, C. M., & Soar, R. S. (1980). How valid are expert opinions about effective teaching? *Phi Delta Kappan, 62,* 131-134.

Crick, J. E., & Brennan, R. L. (1983). *Manual for GENOVA: A GENeralized Analysis Of Variance*. Research and Development Division, American College Testing Program.

Crosnoe, R., Johnson, M. K., & Elder Jr, G. H. (2004). Intergenerational bonding in school: The behavioral and contextual correlates of student-teacher relationships. *Sociology of Education*, *77*(1), 60-81.

Danielson, C. (1997). Enhancing professional practice: A Framework for Teaching. Alexandria, VA: Association for Supervision and Curriculum Development.

Darling-Hammond, L. (2000). Teacher quality and student achievement: A review of state policy evidence. *Education Policy Analysis Archives, 8(1).*

Darling-Hammond, L., & McLaughlin, M. W. (1995). Policies that support professional development in an era of reform. *Phi Delta Kappan, 76*(8), 597-604.

Desimone, L. M., Porter, A. C., Garet, M. S., Yoon, K. S., & Birman, B. F. (2002). Effects of professional development on teachers' instruction: Results from a three-year longitudinal study. *Educational evaluation and policy analysis*, *24*(2), 81-112.

Doyle, W. (1986). Classroom organization and management. *Handbook of research on teaching*, *3*, 392-431.

Emmer, E. T., & Stough, L. M. (2001). Classroom management: A critical part of educational psychology, with implications for teacher education. *Educational Psychologist, 36,* 103–112.

Evertson, C., Anderson, C., Anderson. L., & Brophy, J. (1980). Relationships between classroom behaviors and student outcomes in junior high mathematics and English classes. *American Educational Research Journal*, *17*(1), 43-60.

Evertson, C. M., Emmer, E. T., Sanford, J. P., & Clements, B. S. (1983). Improving classroom management: An experiment in elementary school classrooms. *The Elementary School Journal, 84*(2):173-188.

Fuller, F. F. (1969). Concerns of teachers: A developmental conceptualization. *American Educational Research Journal, 6*, 207-226.

Fuller, F. F., & Bown, O. H. (1975). *Becoming a teacher*. National Society for the Study of Education.

Gage, N. L., & Needles, M. C. (1989). Process–product research on teaching. *Elementary School Journal*, *89*, 253–300.

Gilman, R., & Anderman, E. M. (2006). The relationship between relative levels of motivation and intrapersonal, interpersonal, and academic functioning among older adolescents. *Journal of School Psychology*, *44*(5), 375-391.

Goe, L. (2007). *The link between teacher quality and student outcomes: A research synthesis.* Washington, D.C.: National Comprehensive Center for Teacher Quality.

Good, T. L., & Brophy, J. (2008). *Looking in Classrooms (10th Edition).* New York: Allyn & Bacon.

Good, T. L., & Grouws, D. A. (1977). Teaching Effects: A Process-Product Study in Fourth Grade Mathematics Classrooms. *Journal of Teacher Education*, *28*(3), 49-54.

Graeber, A. O., Newton, K. J., & Chambliss, M. J. (2012). Crossing the borders again: Challenges in comparing quality instruction in mathematics and reading. *Teachers College Record*, *114*(4), 1-30.

Grossman, P. L. (1992). Why models matter: An alternative view on professional growth in teaching. *Review of Educational Research, 62*(2), 171-179.

Grossman, P. L., & Stodolsky, S. S. (1995). Content as context: The role of school subjects in secondary school teaching. *Educational Researcher*, *24*(8), 5-23.

Grossman, P., Loeb, S., Cohen, J., Hammerness, K., Wyckoff, J., Boyd, D., & Lankford, H. (2010). *Measure for measure: The relationship between measures of instructional practice in middle school English language arts and teachers' value-added scores* (No. w16015). National Bureau of Economic Research.

Grossman, P., & Thompson, C. (2008). Learning from curriculum materials: Scaffolds for new teachers? *Teaching and Teacher Education*, *24*(8), 2014-2026.

Hamre, B. K., Pianta, R. C., Mashburn, A. J., & Downer, J. T. (2007). Building a science of classrooms: Application of the CLASS framework in over 4,000 US early childhood and elementary classrooms. *New York, NY, Foundation for Child Development*.

Harris, D. N., & Sass, T. R. (2007). Teacher training, teacher quality, and student achievement. *National Center for the Analysis of Longitudinal Data in Education Research (CALDER). Working Paper*, *3*.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81-112.

Hedges, L. V., & Nowell, A. (1999). Changes in the Black-White gap in achievement test scores: The evidence from nationally representative samples. *Sociology of Education*, *72*, 111-135.

Heneman III, H. G., Milanowski, A., Kimball, S. M., & Odden, A. (2006). Standards-based teacher evaluation as a foundation for knowledge-and-skill-based pay. Consortium for Policy Research in Education Policy Briefs.

Henry, G. T., Bastian, K. C., & Fortner, C. K. (2011). Stayers and leavers early-career teacher effectiveness and attrition. *Educational Researcher*, *40*(6), 271-280.

Hiebert, J. (2003). *Teaching mathematics in seven countries: Results from the TIMSS 1999 video study*. DIANE Publishing.

Hill, H., Kapitula, L., & Umland, K. (2011). A validity argument approach to evaluating teacher value-added scores. *American Education Research Journal, 48*(3), 1-38

Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal, 42*(2), 371-406.

Hobson, A. J., Ashby, P., Malderez, A., & Tomlinson, P. D. (2009). Mentoring beginning teachers: What we know and what we don't. *Teaching and Teacher Education, 25*, 207-216.

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

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

Junker, B. W., Matsumura, L. C., Crosson, A., Wolf, M. K., Levison, A., Wiesberg, J., & Resnick, L. (2006). *Overview of the Instructional Quality Assessment*. (CSE Technical Report #671). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Kagan, D. M. (1992). Professional growth among preservice and beginning teachers. *Review of Educational Research, 62*(2), 129-170.

Kane, T. J., Rockoff, J. E., & Staiger, D. O. (2006). *What does certification tell us about teacher effectiveness? Evidence from New York City*. National Bureau of Economic Research.

Kane, T. J., & Staiger, D. O. (2012). Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains. Policy and Practice Brief. MET Project. *Bill & Melinda Gates Foundation*.

Kane, T., Taylor, E., Tyler, J., & Wooten, A. (2010). *Identifying Effective Classroom Practices Using Student Achievement.* Cambridge, MA: NBER.

Kauffman, D., Moore Johnson, S., Kardos, S., Liu, E., & Peske, H. (2002). "Lost at sea": New teachers' experiences with curriculum and assessment. *The Teachers College Record*, *104*(2), 273-300.

Kennedy, M. M. (2010). Attribution error and the quest for teacher quality. *Educational Researcher*, *39*(8), 591-598.

Kimball, S. M., White, B., Milanowski, A.T., & Borman, G. (2004). Examining the relationship between teacher evaluation and student assessment results in Washoe County. *Peabody Journal of Education , 79*(4), 54-78.

King, F.J., Goodson, L., & Rohani, F. (1998) Higher-order thinking skills: Definitions, strategies, and assessment. Center for Advancement of Learning and Assessment. Tallahassee, FL: Florida State University.

Klem, A. M., & Connell, J. P. (2004). Relationships matter: Linking teacher support to student engagement and achievement. *Journal of School Health*, *74*(7), 262-273.

Klein, A. (2012, June 16). More than half of states now have NCLB waivers. *Education Week.* Retrieved from http://www.edweek.org/ew/articles/2012/07/18/36waivers.h31.html?tkn=ZSMFrarixRk6BCpX%2B9msZwJt%2FKiXlyOAiay0&cmp=clp-edweek

Kolen, M. J., & Brennan, R. L. (1995). *Test equating: Methods and practices.* New York: Springer.

Kulik, J. A., & Kulik, C. L. C. (1988). Timing of feedback and verbal learning. *Review of Educational Research*, *58*(1), 79-97.

Kupermintz, H. (2003). Teacher effects and teacher effectiveness: A validity investigation of the Tennessee value added assessment system. *Educational Evaluation and Policy Analysis, 25,* 287-298.

Lampert, M. (2010). Learning teaching in, from, and for practice: What do we mean?. *Journal of Teacher Education*, *61*(1-2), 21-34.

Lang, J. W., & Kersting, M. (2007). Regular feedback from student ratings of instruction: Do college teachers improve their ratings in the long run? *Instructional Science*, *35*(3), 187-205.

Le Maistre, C., & Paré, A. (2010). Whatever it takes: How beginning teachers learn to survive. *Teaching and Teacher Education*, *26*(3), 559-564.

Little, R. J., & Rubin, D. B. (2002). *Statistical analysis with missing data*. New York: John Wiley & Sons.

Malmberg, L. E., Hagger, H., Burn, K., Mutton, T., & Colls, H. (2010). Observed classroom quality during teacher education and two years of professional practice. *Journal of Educational Psychology*, *102*(4), 916.

Marsh, H. W. (2007). Do university teachers become more effective with experience? A multilevel growth model of students' evaluations of teaching over 13 years. *Journal of Educational Psychology*, *99*(4), 775.

Marzano, R. J. (2004). *Building background knowledge for academic achievement: Research on what works in schools*. Association for Supervision & Curriculum Development.

Marzano, R. J., Marzano, J. S., & Pickering, D. (2003). *Classroom management that works: Research-based strategies for every teacher*. Association for Supervision & Curriculum Development.

Matsumura, L. C., Slater, S. C., Junker, B., Peterson, M., Boston, M., Steele, M., et al. (2006). *Measuring reading comprehension and mathematics instruction in urban middle schools: A pilot study of the Instructional Quality Assessment.* (CSE Technical Report #681). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Matsumura, L. C., Garnier, H., Slater, S. C., & Boston, M. B. (2008). Measuring instructional interactions 'at-scale', *Educational Assessment, 13*(4), 267-300.

Matsumura, L. C., Slater, S. C., & Crosson, A. (2008). Classroom climate, rigorous instruction and curriculum, and students' interactions in urban middle schools. *The Elementary School Journal*, *108*(4), 293-312.

McDonald, F. J., & Elias, P. (1976). The Effects of Teaching Performance on Pupil Learning: Final Report, Volume I, Beginning Teaching Evaluation Study, Phase II, 1974-1976. *Princeton, NJ: Educational Testing Service*.

Meister, D. G., & Melnick, S. A. (2003). National new teacher study: Beginning teachers' concerns. *Action in Teacher Education, 24*(4), 87-94.

Melnick, S. A. & Meister, D. G. (2008).A comparison of beginning and experienced teachers'concerns. *Educational Research Quarterly, 31*(3), 39-56.

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

National Center for Education Statistics. (2002). Dropout rates in the United States 2000. Washington, DC: U.S. Department of Education, Offices of Educational Research and Improvement.

National Commission on Teaching and America's Future. (1996). *What matters most: Teaching for America's future*. Washington, DC: Author.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: Author.

National Council of Teacher Quality. (2014). *2013 State teacher policy yearbook: National summary.*

Newmann, F. M., Lopez, G., & Bryk, A. S. (1998). *The quality of intellectual work in Chicago schools: A baseline report*. Chicago: Consortium on Chicago School Research.

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

Oberski, I., Ford, K., Higgins, S., & Fisher, P. (1999). The importance of relationships in teacher education. *Journal of Education for Teaching: International Research and Pedagogy*, *25*(2), 135-150.

Ost, B. (2009). *How do teachers improve? The relative importance of specific and general human capital* [Electronic version]. Retrieved from Cornell University, School of Industrial and Labor Relations site: http://digitalcommons.ilr.cornell.edu/workingpapers/125/

Papay, J. P. (2012). Refocusing the debate: Assessing the purposes and tools of teacher evaluation. *Harvard Educational Review*, *82*(1), 123–141.

Petersen, N. S., Cook, L. L., & Stocking, M. L. (1983). IRT versus conventional equating methods: A comparative study of scale stability. *Journal of Educational and Behavioral Statistics*, *8*(2), 137-156.

Pianta, R. C., Hamre, B. K., & Mintz, S.L. (2011). *Classroom Assessment Scoring System – Secondary Manual*.

Pianta, R. C., La Paro, K., & Hamre, B. (2004). *Classroom Assessment Scoring System: Pre-kindergarten*. Charlottesville, VA: University of Virginia. Center for Advanced Study of Teaching and Learning.

Piwowar, V., Thiel, F., & Ophardt, D. (2013). Training inservice teachers' competencies in classroom management. A quasi-experimental study with teachers of secondary schools. *Teaching and Teacher Education*, *30,* 1-12.

Porter, A., McMaken, J., Hwang, J., & Yang, R. (2011). Common core standards the new US intended curriculum. *Educational Researcher*, *40*(3), 103-116.

Pourdavood, R. G., Grob, S., Clark, J., & Orr, H. (1999). Discourse and professional growth: processes, relationships, dilemmas, and hope. *School Community Journal*, *9*(1), 33–48.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Reagan, T., Case, K., Case, C. W., & Freiberg, J. A. (1993). Reflecting on "reflective practice": Implications for teacher evaluation. *Journal of Personnel Evaluation in Education*, *6*(3), 263-277.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica, 73,* 2, 417-458.

Rock, D. A., Ekstrom, R. B., Goertz, M. E., & Pollack, J. M. (1985). Determinants of achievement gain in high school. Princeton, NJ: Educational Testing Service.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review, 94,* 2, 247-252.

Roehrig, G. H., & Luft, J. A. (2004). Research Report: Constraints experienced by beginning secondary science teachers in implementing scientific inquiry lessons. *International Journal of Science Education, 26*(1), 3-24.

Roehrig, G. H., & Luft, J. A. (2006). Does one size fit all? The induction experience of beginning science teachers from different teacher-preparation programs. *Journal of Research in Science Teaching, 43*(9), 963-985.

Rogosa, D., Floden, R., & Willett, J. B. (1984). Assessing the stability of teacher behavior. *Journal of Educational Psychology*, *76*(6), 1000.

Rosenshine, B. (1995). Advances in research on instruction. *The Journal of Educational Research*, *88*(5), 262-268.

Ryan, K. (1986). *The induction of new teachers*. Bloomington, IN: Phi Delta Kappa Educational Foundation.

Ryan, A. M., & Patrick, H. (2001). The classroom social environment and changes in adolescents' motivation and engagement during middle school. *American Educational Research Journal*, *38*(2), 437-460.

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

Sass, T., J. Hannaway, X. Zeyu, D. Figlio, & Feng, L. (2010). Value added of teachers in high-poverty schools and lower-poverty schools. CALDER Working paper 52. Washington DC: The Urban Institute.

Schacter, J., & Thum, Y. M. (2004). Paying for high- and low-quality teaching. *Economics of Education Review, 23*(4), 411-430.

Schmidt, W. H., & Houang, R. T. (2012). Curricular coherence and the common core state standards for mathematics. *Educational Researcher, 41*(8), 294-308.

Schoenfeld, A. H. (2002). Making mathematics work for all children: Issues of standards, testing, and equity. *Educational Researcher, 31*(1), 13-25.

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

Shulman, L. S. (1987). Knowledge and teaching. *Harvard Educational Review*, *57*(1), 1-22.

Shulman, J. H., & Colbert, J. A. (1988). *The Intern Teacher Casebook*. ERIC Clearinghouse on Teacher Education, One Dupont Circle, Suite 610, Washington, DC 20036.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford university press.

Neergaard, L., Smith, T., Hochberg, E., & Desimone, L. (2011, September). *Impact of organizational supports for math instruction on the instructional quality of beginning teachers*.Paper presented at the annual conference of the Society for Research on Educational Effectiveness, Washington, DC.

Snow, C. (2002). *Reading for understanding: Toward an R&D program in reading comprehension*. Rand Corporation.

Soar, R. S., & Soar, R. M. (1979). Emotional climate and management. *Research on teaching: Concepts, findings, and implications*, 97-119.

Stanulis, R. N., Little, S., & Wibbens, E. (2012). Intensive mentoring that contributes to change in beginning elementary teachers' learning to lead classroom discussions. *Teaching and Teacher Education, 28*(1), 32-43.

Stein, M. K., Grover, B. W., & Henningsen, M. (1996). Building student capacity for mathematical thinking and reasoning: An analysis of mathematical tasks used in reform classrooms. *American Educational Research Journal, 33*(2), 455-488.

Stein, M. K., & Lane, S. (1996). Instructional tasks and the development of student capacity to think and reason: An analysis of the relationship between teaching and learning in a reform mathematics project. *Educational Research and Evaluation, 2*(1), 50-80.

Stodolsky, S. (1988). *The Subject Matters: Classroom Activity in Math and Social Studies*. Chicago: The University of Chicago Press.

Stodolsky, S. & Grossman, P. (1995). The impact of subject matter on curricular activity: An analysis of five academic subjects. *American Educational Research Journal*, 32, 227–249.

Taylor, E. S., & Tyler, J. H. (2011). *The effect of evaluation on performance: Evidence from longitudinal student achievement data of mid-career teachers*. National Bureau of Economic Research.

Tennessee Department of Education. (2012). Teacher evaluation in Tennessee: A report on year 1 implementation. Nashville, TN: Author.

United States Department of Education. (2009, October 6). U.S. Secretary of Education announces national competition to invest in innovation [Press release]. Retrieved from http://www2.ed.gov/news/pressreleases/2009/10/10062009a.html

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

Valli, L., Croninger, R., & Buese, D. (2012). Studying high-quality teaching in a highly charged policy environment. *Teachers College Record*, *114*(4), 33.

Veenman, S. (1984). Perceived problems of beginning teachers. *Review of Educational Research, 54(2),* 143-178.

Veenman, S. (1987). *On becoming a teacher: An analysis of initial training*. Paper presented at the Conference on Education of the World Basque Congress.

Wenglinsky, H. (2002). The link between teacher classroom practices and student academic performance. *Education Policy Analysis Archives*, *10*(12), 12.

Wang, M. C., Haertel, G. D., & Walberg, H. J. (1997). Learning influences. In H. J. Walberg & G. D. Haertel (Eds.), *Psychology and educational practice* (pp. 199–211). Berkeley, CA: McCatchan.

Wei, R. C. & Pecheone, R. (2010). Teaching performance assessments as summative events educative tools. In Mary Kennedy (ed.), *Teacher assessment and teacher quality: A handbook*. New York: Jossey-Bass.