ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 22-591

Challenges and Tradeoffs of "Good" Teaching: The Pursuit of Multiple Educational Outcomes

David Blazar
University of Maryland College Park

Cynthia Pollard
Brown University

The pursuit of multiple educational outcomes makes teaching a complex craft subject to potential conflicts and competing commitments. Using a dataset in which teachers were randomly assigned to students paired with videotapes of instruction, we both document and unpack such a tradeoff. Upper-elementary teachers who excel at raising students' math test scores often are less successful at improving student-reported engagement in class (and vice versa). Further, the teaching practices that improve math test scores (e.g., cognitively demanding content) can simultaneously decrease engagement. At the same time, paired quantitative and qualitative analyses reveal two areas of practice that support both outcomes: active mathematics with opportunities for hands-on participation; and established routines and procedures to proactively organize the classroom environment. In addition to guiding practice-based teacher education, our mixed-methods analysis can serve as a model for rigorously studying and identifying dimensions of "good" teaching that promote multidimensional student development.

VERSION: June 2022

Suggested citation: Blazar, David, and Cynthia Pollard. (2022). Challenges and Tradeoffs of "Good" Teaching: The Pursuit of Multiple Educational Outcomes. (EdWorkingPaper: 22-591). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/ajht-4d94

Challenges and Tradeoffs of "Good" Teaching: The Pursuit of Multiple Educational Outcomes

David Blazar
University of Maryland College Park
Cynthia Pollard
Harvard Graduate School of Education

June 2022

Abstract

The pursuit of multiple educational outcomes makes teaching a complex craft subject to potential conflicts and competing commitments. Using a dataset in which teachers were randomly assigned to students paired with videotapes of instruction, we both document and unpack such a tradeoff. Upper-elementary teachers who excel at raising students' math test scores often are less successful at improving student-reported engagement in class (and vice versa). Further, the teaching practices that improve math test scores (e.g., cognitively demanding content) can simultaneously decrease engagement. At the same time, paired quantitative and qualitative analyses reveal two areas of practice that support both outcomes: active mathematics with opportunities for hands-on participation; and established routines and procedures to proactively organize the classroom environment. In addition to guiding practice-based teacher education, our mixed-methods analysis can serve as a model for rigorously studying and identifying dimensions of "good" teaching that promote multidimensional student development.

Keywords: instructional practice, teacher effectiveness, math achievement, student engagement, experimental design, mixed-methods research

Correspondence regarding the paper can be sent to David Blazar at dblazar@umd.edu; Department of Teaching and Learning, Policy and Leadership, 2311 Benjamin Building, University of Maryland College of Education, College Park, MD, 20742. We thank staff in the research offices of collaborating districts that provided data; and Casey Archer, Pamela Callahan, Blake Clark, and Monica Anthony at the University of Maryland for research support.

Introduction

Teaching is multidimensional. It pursues multiple educational outcomes and therefore draws on multiple clusters of teachers' knowledge and skills. Over several decades, researchers have defined and documented the importance of varied teacher and teaching competencies, including knowledge of subject matter and how to teach it (Hill et al., 2005); "ambitious" teaching techniques that engage students in challenging tasks to help them make meaning of academic concepts (Spillane & Thompson, 1997); knowledge of student background and regard for student perspectives to create supportive classroom environments (Pianta & Hamre, 2009); and organizational techniques to ensure that lessons are productive and not sidetracked by misbehaviors (Emmer & Stough, 2001). This multidimensional view of teaching aligns with characterizations of students' multidimensional development, encompassing not only knowledge of content but also persistence, self-regulation, and engagement in class activities (Bodovski & Farkas, 2007).

How, then, can teacher educators help teachers excel at or improve in these multiple dimensions? Many researchers advocate for "practice-based" teacher education and professional learning "grounded in…tasks, questions, and problems of practice" (Ball & Cohen, 1999, p. 20), with some further advocating for a "common technical vocabulary" (Lortie, 1975, p. 73; see also Grossman & McDonald, 2008). Theory and a growing research base, however, underscore potential "dilemmas" (Lampert, 2001) and "predicaments" (Cohen, 2011) faced by teacher educators and teachers themselves when pursuing all the competencies described above, and others. Doing so can result in potential conflicts and competing commitments.

Our study contributes to the vast theoretical and empirical literature on teaching practice—and its inherent dilemmas and predicaments—in three unique ways. The first contribution is methodological. Like others (Ball & Forzani, 2009; Grossman et al., 2009; McDonald et al, 2013), we argue that building towards a common technical vocabulary requires systematic empirical validation that bridges the strengths of scholars who study teachers and teaching from varied disciplinary frames. Our sequential explanatory mixed-methods design (Ivankova et al., 2006) begins with a tradition often pursued by economists to causally link teachers, their characteristics and practices, and student outcomes (e.g., Hanushek, 2002). We do so by using a dataset in which teachers were randomly assigned to class rosters within schools—a rarity in education research, but also a key design strategy for drawing robust causal claims (Charalambous & Delaney, 2020). Next, we build on the tradition of teacher education researchers to directly observe classroom instruction (e.g., Ball et al., 2009). Drawing from established observation protocols and more open-ended qualitative inquiry, we consider a broad set of teaching practices that may be responsible for improved student outcomes. Some scholars have identified tensions between these traditions and frameworks, noting that "quality" teaching as defined by classroom observation research may not be the same practices as those that are "successful" or "effective" at improving student outcomes (Fenstermacher & Richardson, 2005). Rather than setting one against the other, we recognize that both frameworks can bring insight to the study of "good" teaching and the tradeoffs teachers may face in their work with students. Our use of the term "good" aims to sit at the intersection of these perspectives.

The second contribution of our paper is substantive, where we provide rigorous quantitative documentation on some of the dilemmas and predicaments of teaching that have been conceptualized by others. We find that, on average, upper-elementary teachers who excel at improving students' test-based achievement in mathematics often are less successful at increasing student-reported engagement in math classroom activities (and vice versa) (see also Blazar, 2018; Blazar & Kraft, 2017). The pattern holds not just when examining student achievement on high-stakes assessments—sometimes problematized by teacher education scholars for being misaligned with goals of instructional reform (e.g., Fenstermacher & Richardson, 2005)—but also when using an assessment designed to capture students' conceptual knowledge of math content. More worrisome, when we link classroom observation scores to these same student outcomes, we find that teaching practices that result in increased math test scores (e.g., cognitively demanding, error-free content) can simultaneously result in *decreased* student engagement. While our study is interested in a broad set of teaching practices, we recognize that this pattern counters longstanding reform orientations—particularly in mathematics—towards ambitious teaching specifically (Cobb & Jackson, 2011; Spillane & Thompson, 1997).

Our third contribution aims to make sense of these tradeoffs, ultimately providing direction for the field on a "common technical vocabulary" of practices that can improve *both* students' test-score performance and classroom engagement. To do so, we conduct fine-grained qualitative analyses of the mathematics lessons of teachers who excel at improving both students' test scores and self-reported classroom engagement, versus the lessons of teachers who improve test scores only. These analyses point to benefits of teaching practices in two key areas. The first is active mathematics, in which teachers provide opportunities for hands-on participation, physical movement, or peer interaction. These activities overlap with ambitious teaching techniques that often make use of manipulatives and tactile activities in the service of building conceptual understanding. However, our study suggests that the active component—and not necessarily the push for conceptual understanding—supports student test-score performance and engagement.

The second area of practice that we find supports both sets of student outcomes is established practices and routines, where teachers communicate their expectations for students in a way that increases efficiency and order in the classroom. In our discussion, we speak to tensions around classroom and behavior management, which often are described as exclusionary (Milner & Tenore, 2010) and are challenged by some scholars for reducing the complexity of teaching to a simple set of moves (McDonald et al., 2013). At the same time, our findings suggest that when routines are proactive rather than reactive, they can create classroom conditions that benefit students' test-score performance and engagement.

Ultimately, we argue that these two areas of teaching practice should serve as foundational components of practice-based teacher education and professional learning. We also advocate for continued empirical validation and identification of "good" teaching practices by combining the strengths of experimental designs with qualitative observations of classroom interactions.

Interdisciplinary Perspectives on "Good" Teachers and "Good" Teaching

The central question and challenge that motivates this paper is a longstanding one: What makes for "good" teaching? For decades, researchers from multiple disciplinary traditions have weighed this question, using theory and empirical evidence to identify characteristics of teachers and instructional practices that best serve their students. However, the approaches scholars take differ substantially, from fine-grained qualitative examination of classrooms and teacher-student interactions, on one hand, to quantitatively linking teachers to student outcomes in large-scale datasets, on the other. Pianta and Hamre (2009) state this tension concretely when they note that:

> "…studies of student achievement gains have been important in laying a foundation for inquiry into classroom effects…[but] they fail to articulate specific processes that may lead to student learning and positive social adjustment. The problem with this atheoretical approach…is reflected in Hanushek's (2002) definition of teacher quality, `Good teachers are ones who get large gains in student achievement for their classes; bad teachers are just the opposite' (p. 3); this definition and much of the research using the value-added paradigm [that links teachers to gains in student outcomes]…provide only limited guidance to efforts to improve teaching and teacher education…in the sense that they do not inform how training and professional development might focus attention or shape teacher behavior." (p. 110)

In some ways, our study builds closely from this perspective, particularly in the push for theory-based empirical analysis. Ultimately, we argue that "good" teach*ers* are able to improve student outcomes because they engage in "good" teach*ing* practices. At the same time, we are much more optimistic about how traditions can complement each other, particularly when faced with challenging questions.

Deeply connected to this first question, then, is a second one: Why is "good" teaching such hard work? Our review of the literature on this question suggests that scholars who think about teachers and teaching from distinct traditions actually have a fair amount in common. From the perspective of an educator wrestling with her own and others' practice, Lampert (2001) writes that, "One reason that teaching is a complex practice is that many of the problems a teacher must address to get students to learn occur simultaneously, not one after another" (p. 2). Teachers are charged with delivering content while also supporting social-emotional development, and they must address the needs of the individual and of the group. This balancing act occurs over and over again during the span of a lesson, a school day, and the school year. Further illustrating this point, Pianta and Hamre (2009) synthesize a vast amount of scholarship on teaching and student development to conceptualize several domains of teachers' work: (i) emotional supports that attend to students' sense of attachment (Ainsworth et al., 1978) and self-determination (Connell & Wellborn, 1991); (ii) organizational supports that help students build self-regulatory skills for behavior, time-use, and attention in the classroom (Emmer & Stough, 2003; Paris & Paris, 2003); and (iii) instructional supports, in which teachers make use of curricula and learning activities to support students' metacognitive skills that are critical to academic development (Veenman et al., 2005). Each component is complex in isolation, and even more complex as teachers pursue these components simultaneously.

The policy circumstances and reforms that guide teachers' work create additional, often competing commitments (e.g., Cohen, 2011). Following the tradition of Hanushek (2002) and accountability-oriented policy, teachers are asked to improve student test scores on high-stakes assessment. But they also enact classroom-based policies in the form of new curricula and standards, which may or may not align with testing regimes. For example, in mathematics—which is the content focus of this study—educators and scholars often have advocated for inquiry-oriented or ambitious practices that elicit and build on students' mathematical thinking in order to make meaning of mathematical concepts. Professional boards and panels including the National Council of Teachers of Mathematics (NCTM; 2014) the National Mathematics Advisory Panel (2008), and the National Governors Association (NGA; 2010) recommend that teachers support students by: providing explanations for mathematical phenomena (Leinhardt, 2001); drawing explicit links between multiple mathematical representations (Ball et al., 2005); comparing multiple strategies to solve a single problem (Rittle-Johnson & Star, 2007); and examining patterns to make generalizations (Callejo & Zapatera, 2017). The constructivist-oriented push to organize teaching and learning around sensemaking also is associated with calls for student-centered as opposed to teacher-led instruction (Jones, 2007). At the same time, the goals of ambitious and student-oriented inquiry practices can be difficult to achieve in the face of top-down pushes for testing and accountability (Cohen & Hill, 2008).

The inherent challenge of teaching towards multiple, potentially competing goals is further documented in the work of economists seeking to monitor, measure, and reward performance. For example, Holmstrom and Milgrom (1991) describe the multitask nature of teachers' work as a potential "problem" because of multiple incentives that, on one hand, pursue basic skills and, on the other hand, strive for higher-level thinking and curiosity. Economic analyses also provide strong causal support for the phenomena described thus far. Experimental studies are not the only way to explore the nature of teachers' work, but they do address concerns that teachers' interactions with students and their impacts on varied student outcomes can be shaped by the purposeful matching of teachers and students within classrooms (Steinberg & Garrett, 2016). To address this concern, Kraft (2019) used a dataset in which teachers were randomly assigned to class rosters within schools and found that teachers have large causal effects on students' achievement on complex mathematical tasks, as well as on their grit, effort, and growth mindset in math. However, teachers' effects on these varied student outcomes were only weakly correlated at best. In similar analyses, Blazar (2018) and Blazar and Kraft (2017) used a subset of the data described in this paper to examine teachers' contributions to students' mathematics test scores versus their contributions to students-reported engagement in classroom activities, finding negative correlations as large as -0.4. (These correlations are replicated below.) These findings suggest that, on average, improvements in students' test-based achievement in math come at the expense of improvements in classroom engagement (and vice versa).

How, then, can teachers and teacher educators begin to unravel these challenges, tensions, and tradeoffs? Circling back to Pianta and Hamre's (2009) critique of the value-added paradigm, we agree that causal inquiry of classroom and teacher effects is a critical foundation, but that simply documenting null or negative correlations between teachers' effects on varied student outcomes is incomplete. Without attention paid to specific classroom features, the mechanisms—and thus, potential solutions—for the tradeoff is unclear. Is it that teachers follow goals for ambitious teaching and that these efforts are engaging for students, while test-based achievement is not measuring conceptual understanding that ambitious teaching aims for? Conversely, is it that a focus on test-based achievement pushes teachers to engage in "drill and kill" test preparation, which is less engaging for students (Kolluri, 2018)? It also is possible that teachers do focus on student sensemaking activities that have long been the focus of reformers; but, environments where students do much of the "heavy lifting" may feel disordered, perhaps disrupting students' relationship to the teacher and their sense of belonging and engagement. Without downplaying the reasons why many individuals advocate for inquiry-oriented instruction, some teacher educators (Lampert, 2001; Kennedy, 2005) and developmental psychologists (Kirschner et al., 2006) have questioned whether students themselves enjoy taking on more responsibility for their learning and higher expectations for conceptual understanding; some students may prefer clear routines and procedures delivered by teachers.

In turn, our analyses aim to identify a set of "good" teaching practices that can simultaneously achieve multiple goals, in the form of multiple student outcomes. Our study is not alone in this endeavor. Efforts to codify teacher-student interactions in validated and widely used classroom observation protocols—including two used in this study—and then link these practices to student outcomes guide much of the work around "core" or "high-leverage" practices (Ball et al., 2009;

Forzani, 2014; Grossman et al., 2009; McDonald et al, 2013). A recent review of the literature on this topic provides some evidence of links between core teaching practices and student outcomes (Charalambous & Delaney, 2020), including an orderly environment (Bell et al., 2012), time on task (Stronge et al., 2011), and the cognitive and disciplinary demand of instruction (Blazar, 2015). At the same time, the authors of the review describe a need for "stronger and more systematic empirical validation" from studies that use causal research designs and examine a broad set of teaching practices and a broad set of student outcomes within a single study (Charalambous & Delaney, 2020, p. 356).

Variation in research design quality also has led to mixed conclusions about the benefits of specific teaching practices and formats over others. For example, longstanding debates on direct versus student-centered instruction remain unresolved due in part to research design. In a recent meta-analysis on this topic that pooled results from all quantitative analyses no matter the design, the authors concluded benefits of direct instruction to a range of student outcomes; yet, the much smaller subset of random assignment analyses often returned much smaller or null effects (Stockard et al., 2018).

In other words, to better understand challenges and tradeoffs of "good" teaching, we must build on the strengths of multiple disciplinary traditions. Our study fills this gap.

Sample and Experimental Design

In order to provide guidance on how teachers may simultaneously support students' academic achievement and classroom engagement, we leverage a unique dataset from the National Center for Teacher Effectiveness (NCTE) collected over a three-year period (2010-11 through 2012-13).[1] The

[1] Since the NCTE data were collected between the 2010-11 and 2012-13 school years, advancements have been made in "core" and "high-leverage" practices, including providing resources to teachers. (See, for example, https://library.teachingworks.org/curriculum-resources/high-leverage-practices/.) Therefore, it is possible that teachers have grown more sophisticated in their understanding and implementation of these types of strategies over the past decade. While we cannot directly test this claim, we acknowledge that such trends may impact our findings if average practices differ today versus when the NCTE data were collected. That said, the research design and scope of the NCTE study is fairly rare in the study of teaching, and so serves as a valuable resource. Further, because the data are older, we are able to examine the predictive power between measures collected from upper-elementary students from that time to high school outcomes captured several years later; these analyses provide empirical justification for focusing on the student measures (e.g., test-based achievement in mathematics, student-reported engagement) that ground our work.

full project sample includes fourth- and fifth-grade teachers and their students in four school districts on the east coast of the United States (U.S.). While our analyses focus on teachers' mathematics lessons, teachers were generalists who taught all subject areas. A key feature of these data is that, in the spring of 2012, the NCTE project team worked with staff at participating schools to randomly assign a subset of teachers (n = 53) to class rosters (n = 829 students) in the same school and grade level (i.e., randomization blocks) that were constructed by principals or other school leaders.[2] Random assignment allows us to estimate the contribution of teachers to student outcomes that is not confounded by the preferences of teachers, school leaders, students, and families.

In Table 1, we show that the subset of teachers in the experimental portion of the study look similar to those in the full NCTE dataset (n = 309 teachers 9,141 students) on a host of characteristics including education, experience, and content knowledge (p = 0.938 on a joint test of significance comparing the characteristics of teachers in the two groups).[3] Like many districts across the U.S., the vast majority of participating teachers were female (84%) and White (67%). Characteristics of students also match those of urban school districts, with variation in terms of race/ethnicity (e.g., 37% African American; 23% Hispanic) and socioeconomic status (67% eligible for free or reduced-price lunch). We observe some statistically significant differences in student characteristics across samples, though the magnitudes of these differences tend to be small.

In Table 2, we provide estimates to confirm the success of the randomization process at creating balanced groups. In a traditional experiment, one can examine balance at baseline by

[2] To increase internal validity and to achieve our project aims, we exclude full randomization blocks where: (i) teachers dropped out of the experiment before the start of the 2012-13 school year (six randomization blocks with nine teachers); (ii) 50% or more of students switched out of their randomly assigned teachers' classroom (three randomization blocks with eight teachers and 221 students); or (iii) teachers were missing classroom observations (two randomization blocks with five teachers and 127 students). Because each randomization block is equivalent to its own experiment, dropping individual ones does not threaten the internal validity of results.

[3] Background teacher characteristics come from a project-administered survey that included a math assessment of items from the Mathematical Knowledge for Teaching assessment (Hill et al., 2008) and the Massachusetts Test for Educator Licensure. Background student characteristics come from districts' administrative records.

calculating differences in average characteristics of participants between the treatment and control groups. In this context, though, treatment consisted of multiple possible teachers within a given randomization block. (Among the 24 total randomization blocks, 19 had two teachers and five had three teachers.) Thus, to examine balance we examined the relationship between the assigned teacher's predicted effectiveness at improving students' math test scores in years prior to the experiment and baseline student characteristics, controlling for randomization block.[4] Of the baseline student characteristics, only students' eligibility for special education services is related to baseline teacher effectiveness. When we test that all the baseline characteristics jointly predict baseline teacher effectiveness, we cannot reject the null hypothesis of no difference ($p = 0.583$).

A second possible threat to our ability to causally link teachers and their practices to student outcomes is attrition due to non-compliance amongst participating students. Of the 829 students in the experimental sample, 163 (20%) switched out of their randomly assigned teachers' classroom before the start of the school year for several reasons including moving to a different school or district (see Appendix Table 1, and Blazar, 2018 for additional details). All analyses presented below exclude these students, because they no longer were part of primary data collection. The compliance rate in this study (80%) is substantially higher than other similar experiments in which teachers were randomly assigned to classes (i.e., Kane et al., 2013). However, non-compliance still could bias results if it were a result of the effectiveness of students' randomly assigned teacher. We find that this is not the case. The average difference between compliers and non-compliers in terms of the baseline effectiveness of their randomly assigned teacher is small (0.007 SD) and not statistically significantly different from zero ($p = 0.222$).

---

[4] We focus on teachers' effects on math test scores as the baseline measure of teacher effectiveness because this a measure that can be purged of sorting bias without an experimental design (Kane et al., 2013). Other dimensions of teacher effectiveness that are of interest in this paper (i.e., teachers' effects on student engagement) retain bias under non-experimental conditions (Blazar, 2018).

Data for Quantitative and Qualitative Inquiry

The data used in this study include students' math test scores and self-reported classroom engagement, videotaped lessons of teachers' math instruction, and a teacher survey capturing their engagement in test-preparation activities. The student outcome data are used in the quantitative analyses, as well as in selection of the qualitative sample of teachers who differ in their effectiveness at improving these outcomes. Videotaped lessons also are used in both types of analyses.

Student Outcomes

Math assessments. The NCTE data include two types of math assessments. One was developed by the NCTE research team with the Educational Testing Service, designed to capture students' understanding of upper-elementary mathematics topics: numbers and operations, algebra, geometry, and measurement. Lynch et al. (2017) coded items for their format and cognitive demand and found that many asked students to solve non-routine problems, including looking for patterns and explaining their reasoning; 20% of items required short responses. Internal consistency reliability (α) is 0.82 or higher for each test form (Hickman et al., 2012; see Table 3 for univariate and pairwise descriptive statistics). The other type of math assessment is state tests. In the coding analyses, Lynch et al. (2017) found that, in two districts from the same state, the test was closely aligned to the sorts of conceptually oriented tasks emphasized in the project assessment (r = 0.81 across the state and project assessments). A third district's assessment included a combination of conceptual and procedural items, while the fourth district's assessment focused primarily on procedures (r = 0.73 and 0.68, relating scores on the two types of math assessments for these two districts, respectively). Triangulating results across two math assessments allows us to consider whether links between teachers' practices and student performance are driven by state-specific test designs, or by teachers' engagement in test-preparation activities aligned to state tests but not the project test.

For both math assessments, test scores were available at the end of the experimental year and at baseline, all of which we scaled to have a mean of 0 and standard deviation (SD) of 1. Of the students in our final analysis sample, just three were missing end-of-year state tests and 38 were missing the project test. Missingness was not related to baseline teacher effectiveness (see Appendix Table 1). Therefore, to maximize our sample size, for the state tests we imputed missing scores to the mean of students' randomization block; for the project test, we imputed missing scores using the state assessments, given a high correlation between the two (see Table 3).

Student surveys. The NCTE project staff administered a survey to students at the end of each school year to capture self-reported social skills, with Likert-scale items adapted from other large-scale research projects including Tripod. We focus on one construct that emerged from these items: *Engagement and Happiness in Class* (5 items; $\alpha = 0.76$; see Table 3 for item text and Blazar & Kraft, 2017 for exploratory factor analyses). To create final scores, we averaged across all items and then standardized to a mean of 0 and SD of 1. As with the math assessments, few students with missing *Engagement* scores, and missingness was not related to baseline teacher effectiveness (see Appendix Table 1); we imputed missing scores to the mean of students' randomization block.

Students may have responded to the *Engagement* items in varied ways, focusing on their engagement in class and math activities specifically or their enjoyment of and happiness in the classroom environment more generally. For example, watching videos may be enjoyable for students but not grounded in core content. To address this possibility, in Table 4, we provide evidence of predictive validity by relating *Engagement and Happiness in Class* and math test scores from upper-elementary school to later test scores and to the Student Engagement Instrument (SEI; Betts et al., 2010) in high school. Longer-run test scores were available from all four districts, while SEI scores came from one. We focus on two dimensions from the SEI that align closely with the *Engagement* measure: *Control and Relevance of School Work* ($\alpha = 0.8$) and *Future Goals and Aspirations* ($\alpha = 0.78$).

Student-reported *Engagement and Happiness in Class* from upper-elementary school predicts all three outcomes in high school, with the strongest relationships to the two SEI survey measures (as expected). These patterns indicate that the *Engagement* measure—as well as the two math test scores—provide valuable information about students' future trajectories in school.

Teaching Practices

Videotaped lessons. Teachers contributed three videotapes of their mathematics lessons each year of the study, with an average of seven total videos per teacher. Capture occurred with a freestanding, three-camera, digital recording device and lasted between 45 and 60 minutes. One camera focused on the front of the classroom, while the other two cameras focused on student tables and desks. One microphone attached to the recording device and another worn by the teacher picked up student talk from around the classroom. Teachers were allowed to choose the dates for capture in advance and were directed to select typical lessons and exclude days during which students were taking a test.[5] Each videotaped lesson is accompanied by a transcript.

In addition to examining the videotaped lessons for qualitative inquiry, our quantitative analyses rely on scores generated by trained raters on two established observation instruments: the Mathematical Quality of Instruction (MQI; Learning Mathematics for Teaching Project, 2011) and the Classroom Assessment Scoring System (CLASS; Pianta, La Paro, & Hamre, 2008). Following protocols from instrument developers, two raters scored lessons on the MQI, and one rater scored lessons on the CLASS. Each rater watched lessons in segments (7.5 minutes for the MQI and 15 minutes for the CLASS), and provided scores on several items that fall under broader dimensions of instruction that we describe next and that are justified based on theory, as well as exploratory and

[5] While it is possible that teachers purposeful selected specific lessons for observation, teachers did not have any incentive to select lessons strategically as no rewards or sanctions were involved with data collection or analyses. Analyses from separate projects also indicate that teachers are ranked almost identically when they choose lessons themselves compared to when lessons are chosen for them (Ho & Kane, 2013).

confirmatory factor analyses (Blazar et al., 2017; see Appendix Table 2 for item text). Teacher-level scores average across raters (where applicable), segments, items, and lessons.[6]

The MQI captures both instructional formats and instructional practices. In terms of formats, *Teacher-Led Instruction* identifies the amount of time teachers spend leading discussion or presentation of content, relative to time in *Small Group, Partner, or Independent Work*. Raters can score instruction as both formats, which generally capture instances in which students work in centers and teachers lead instruction with one group at a time. These formats are mutually exclusive and scored 0 = "no" or 1 = "yes". In our sample, roughly 47% of instruction was teacher-led, while a third was small-group, partner, or independent work, and 20% was a combination of both (see Table 3). *Whole-Class Discussion* is another mode of instruction—not mutually exclusive with the others—capturing instances in which teachers are in charge of the class but where they facilitate discussion of students sharing their thinking, explaining their reasoning, and building on one another's contributions. On a 1 to 3 quantity scale from "none" to "most/all", average teacher scores are just slightly higher than 1.

The remaining dimensions are quality rather than quantity codes. Also from the MQI, *Ambitious Mathematics Instruction* (10 items, scored 1 = "not present" to 3 = "present and sustained") identifies the complexity of the tasks that teachers provide to their students and their interactions around that content, including explicit linking between multiple mathematical representations and teachers' facility responding to student ideas. The quality of ambitious teaching is positively correlated with the amount of time teachers spent on discussion (r = 0.38), but not with time spent in small group, partner, or independent work—even though ambitious activities can occur in teacher-student or student-student interactions. Like for the *Whole-Class Discussion* code, *Ambitious Instruction* was

---

[6] To account for variation in the number of lessons teachers contributed to the dataset, we calculated predicted, shrunken teacher-level scores using the following multilevel model:

$$OBSERVATION_{ljt} = \gamma_j + \varepsilon_{ljt}$$

where the outcome is the observation score for lesson *l* from teacher *j*. Teacher-level random effects, γ_j, are our parameters of interest, which capture mean scores for each teacher shrunken back to the sample mean based on the number of lessons.

relatively rare (average score = 1.26 out of 3). *Mathematical Errors* (3 items, with same scale as above) identifies any mathematical errors or imprecisions that the teachers introduce into the lesson that go uncorrected. Thus, unlike for the other dimensions of practice, higher scores reflect worse performance. Errors were more common for teachers who spent more time leading instruction (r = 0.21) and less common for teachers who spent more time having students work in small groups (r = -0.12) or leading instruction in small groups (r = -0.16). Teachers who made more errors tended to engage less in ambitious teaching (r = -0.26), capturing in part the fact that the MQI does not allow instructional activities to be coded as ambitious if they are mathematically incorrect.

From the CLASS, *Emotional Support* (3 items, scored 1 = "low" to 7 = "high") focuses on teachers' interpersonal relationships with students, including the extent to which teachers create a positive classroom climate and teachers' respect for student ideas. *Classroom Organization* (3 items, with same scale as above) captures teachers' behavior management skills and the pacing of the lesson. The CLASS also includes a single item, *Student Engagement*, that instrument developers separate from the other dimensions. We include this item in some of our analyses because of its close connection to student-reported *Engagement*. Scores from the CLASS are moderately to strongly correlated with each other (r =0.47 to 0.61). However, we keep them as separate because the dimensions have distinct theoretical underpinnings (Pianta & Hamre, 2009) and factor analyses identify them as unique (Blazar et al., 2017). Correlations with scores from the MQI—both formats and practices—are no higher than 0.23. As described earlier, the CLASS includes another dimension of practice on *Instructional Supports*, which we exclude from our quantitative analyses given theoretical and empirical overlap with the instructional components from the MQI (Blazar et al., 2017).

Psychometric analyses indicate that rater scores on each of the dimensions described above adequately capture the underlying constructs of interest. For the MQI, where two raters scored each lesson, inter-rater agreement rates are 0.79 for format of instruction (e.g., teacher-led versus small group), 0.92 for *Whole-Class Discussion*, 0.74 for *Ambitious Mathematics Instruction* and 0.86 for *Mathematical Errors* (see Table 3). We also calculated adjusted intra-class correlations (ICCs) that capture the amount of variation at the teacher level. For the dimensions of practice scored on a quality scale, estimates are 0.74 for *Ambitious Mathematics Instruction,* 0.56 for *Mathematical Errors,* 0.53 for *Emotional Support,* 0.63 for *Classroom Organization*, and 0.28 for the single *Student Engagement* item. For the instructional formats and modes scored on a quantity scale, estimates are 0.66 for *Teacher-Led Instruction*, 0.36 for *Small Group, Partner, or Independent Work*, 0.54 for combined formats, and 0.47 for *Whole-Class Discussion*. Like with *Student Engagement*, these estimates are lower than for other dimensions likely due to the fact that they are single items; more items generally helps to increase reliability. These reliability estimates are consistent with those generated from similar studies (Bell et al., 2012; Kane et al., 2013). We standardized all teacher-level observation scores to have a mean of 0 and SD of 1.

Teacher survey capturing test-preparation activities. Finally, at the end of each school year, teachers completed a survey capturing time spent on five test-preparation activities: using standardized test items in instruction, incorporating item formats, teaching test-taking strategies such as process of elimination, reviewing concepts most likely to be found on the state test, or focusing instruction on students expected to score just below a given performance level on the state test (1 = "never or rarely" to 4 = "daily"; $\alpha = 0.80$). We averaged Likert-scale scores across items and then standardized the composite to have a mean of 0 and SD of 1. Aligned to theory, procedurally based test preparation is associated with lower *Ambitious Mathematics Instruction* ($r = -0.22$; see Table 3).

Analyses

Estimating and Correlating Teachers' Effects on Math Test Scores versus Engagement

Both of our quantitative and qualitative analyses rely first on identifying teachers' contributions to students' math test scores versus student-reported *Engagement and Happiness in Class.* The randomized design allows for a straightforward approach to estimate these teacher effects:

$$OUTCOME_{isgjt} = \nu_{sg} + \delta X_{it-1} + \varphi \bar{X}^{j}_{it-1} + (\tau_j + \varepsilon_{sgjt}) \quad (1)$$

We use $OUTCOME$ to refer to the two math assessments or *Engagement and Happiness in Class* for student *i* in school *s* and grade *g*, working with teacher *j* at time *t* (i.e., the end of the random assignment school year). To match the experimental design, we control for fixed effects for randomization block (i.e., school-grade combinations), ν_{sg}, and class characteristics ($\bar{X}^{j}_{it-1}$) that average baseline student-level characteristics, including demographics and prior-year test scores, to the class level. We also control for baseline student characteristics, X_{it-1}, to increase the precision of our estimates.

In our model, τ_j is our teacher effect estimate. Because the randomized design successfully created balanced groups, we can be assured that these estimates capture the contribution of teachers to student outcomes that is not confounded with other factors, beyond those already controlled for in the model. We shrink the teacher effect estimates back toward the mean based on their precision by fitting a random effects model (Koedel et al., 2015). This approach is beneficial for minimizing error in the teacher effects estimates and, in turn, mitigating attenuation in correlations between teachers' effects on varied student outcomes.

Quantitative Analyses Linking Teaching Practices to Student Outcomes

Second, we quantitatively examined whether certain dimensions of teachers' classroom practice result in improved student math test scores versus classroom engagement. To do so, we specified versions of equation (1) that predicted students' math scores or their *Engagement and Happiness in Class* as a function of instructional quality scores from the MQI and CLASS instruments, as well as our test-preparation composite.

Although teachers were randomly assigned to classes, teaching practices were not randomly assigned to teachers. This means that we could estimate the causal effect of students' *exposure* to teachers and their practices but not the causal effect of the practices themselves. To address this limitation, our preferred models are conditional ones in which we include all teaching practices as independent variables in the same model. This approach aims to isolate the effect of one teaching practice on student outcomes that is not confounded with another. Due to moderate to strong correlations between some teaching practices (particularly on the CLASS instrument; see Table 3), we also estimate unconditional models that added these practices into separate regression models, which we show in an appendix. Another concern when linking teaching practices to student outcomes is that the same student behaviors may show up on both the left- and right-hand side of our regression models, inflating our primary estimates of interest. Therefore, we use out-of-year observation and test-preparation scores (see Kane et al., 2011 for a similar approach).[7]

Qualitative Observation of Mathematics Lessons

Following our sequential explanatory mixed-methods design (Ivankova et al., 2006), we paired our quantitative results with qualitative analysis looking more in-depth at the mathematics lessons of a subset of teachers. The primary goal of our qualitative work was to identify teaching practices that may simultaneously support students' mathematics achievement and classroom engagement. Qualitative analyses also help illustrate, expand, and elaborate patterns that emerged from the quantitative results (Hill et al., 2008), and afford open-ended exploration of the relationship between improved math test scores versus student engagement, with the intent of developing themes from the data that were not visible in the quantitative results (Creswell, 2003).

To achieve this goal, we first randomly selected teachers (*n* = 12) who were successful at raising students' math test scores (i.e., top tercile) on both the state and project math assessments, but varied in the extent to which they made students engaged and happy in class (i.e., six teachers each from top and bottom terciles). It was important to hold constant teacher effects on one of these two types of

[7] Two teachers only had observation scores in the current school year, and so we use these scores. In Appendix Table 1, we show that these teachers do not differ from others in terms of baseline estimates of their contributions to students' math performance.

student outcomes; otherwise, qualitative observation could not disentangle classroom features associated with one versus the other. We opted to hold constant teachers' effects on math scores and exploit variation in teachers' effects on student engagement because—as we describe below—the quantitative analyses were more puzzling when linking classroom observation scores to the engagement measure, relative to math test scores. By randomly selecting a subset of teachers, we aimed to capture typical instructional practice within each of the two groups, while also making the task of observing multiple lessons per teacher feasible. We followed the procedure used in a prior mixed-methods study of the same data in which six teachers per group was sufficient to identify themes and cross-group differences (Blazar et al., 2016).

Next, we randomly selected three lessons per teacher, which is the number of lessons needed to achieve sufficiently high reliability through observation (Hill et al., 2012). Then, we randomly assigned raters to lessons, ensuring that all raters were assigned at least one lesson per teacher. In total, our team included six raters: the two authors, who are trained and certified as raters on either the MQI or CLASS instrument, and four research assistants who are former K-12 classroom teachers with experience observing and rating instruction.[8] All raters were blind to whether teachers were in the top or bottom tercile of effects on student engagement.

While viewing lessons, raters independently noted salient elements of instruction that may be related to student engagement by annotating lesson transcripts. The observation protocol purposefully was unstructured and open-ended to allow for new and potentially unexpected themes to emerge (Creswell, 2003). After reviewing all lessons for a given teacher, the full research team met as a group to discuss notes and to identify features of classroom instruction or environment that were evidenced

[8] Due to language in the NCTE teacher consent forms, it only was possible for the two authors to watch the original videotaped lessons, given our role as researchers on the original project. The additional raters read lesson transcripts, allowed under current data sharing agreements. Therefore, we made sure that one of authors and two to three additional raters were assigned to each lesson.

in multiple lessons of the focal teacher. Following Krueger (1994) the authors served as moderator and assistant moderator during these group discussions, posing questions, prompting raters to speak, and encouraging all raters to participate. One rater was randomly assigned to write a short memo describing these patterns and synthesizing the group conversation.

After repeating this process for all 12 teachers, the authors undertook a data-driven approach to thematic analysis of the memos (Creswell, 2003): we first segmented data into smaller chunks and tagged text, noting emergent trends and patterns. We then looked across these patterns and refined them into codes. Refining the codebook often involved reviewing original lesson transcripts and videos which provided additional clarity and examples. We then revisited memos to apply the codes. To be identified as a common theme, codes had to apply to multiple lessons per teacher and to multiple teachers. Finally, we noted whether themes differentiated low- versus high-engagement teachers, all of whom were successful at improving students' math performance. These themes can provide guidance in terms of the teaching practices that support both students' academic performance and their classroom engagement.

Results

Tradeoffs Between Teachers' Effects on Students' Math Test Scores versus Engagement

In Table 5, we examine correlations between upper-elementary teachers' effects on students' test scores on the two math assessments and their effects on students' *Engagement and Happiness in Class*. In Panel A of Table 5, we include all 53 teachers, while in Panel B we exclude bivariate outliers based on Cook's D (Cook, 1977).[9] Cells that present correlations between teacher effects on students' test scores versus engagement are highlighted in gray; the other cells show correlations between teacher effects on the two different math assessments.

[9] More specifically, we exclude teachers whose value of Cook's D is greater than $4/(n - k - 1)$, where n is the sample size, k is the number of predictors (in our case, just one), and 1 is a degrees of freedom correction. This approach is recommended, in particular, for small sample sizes (Kutner et al., 2005).

Consistent with prior work (Blazar, 2018; Blazar & Kraft, 2017), we find that teachers' effects on students' math test scores consistently are negatively correlated with teachers' effects on students' *Engagement and Happiness in Class*. Correlation coefficients range from -0.24 to -0.43, and all are statistically significantly different from zero. The negative correlations do not appear to be driven by the specific math assessment used. Correlations also are similar when we exclude outliers. We provide a visual illustration of these relationships in Figure 1, with x- and y-axes scaled to [-1,1] so that the slopes of the lines match the correlation coefficients from Panel A of Table 5 that includes the full sample of teachers. Each circle or hollow triangle represents an individual teacher. Circles and the solid best-fit line represent the relationship between teachers' effects on student *Engagement* versus the state math test, while triangles and the dashed best-fit line represent the relationships teachers' effects on *Engagement* versus the project math test. The top horizontal band (i.e., top tercile for teachers' effects on students' math test scores) and the right- and left-most corners (i.e., top versus bottom terciles for teachers' effects on student engagement) captures the set of teachers from which we randomly selected our sample for qualitative analysis.

While the absolute values of the correlations shown in Table 5 and Figure 1 are not "strong" relative to some thresholds, we argue that a statistically significant and negative correlation is substantively meaningful no matter the magnitude. Negative correlations indicate that, on average, teachers who improve test-score performance do so at the expense at student engagement and happiness in class (and vice versa). Further, the classroom practices and activities that go into improving test scores may be quite different from those that improve student engagement. We explore this possibility in the next set of results.

Tradeoffs Between Teaching Practices that Benefit Math Test Scores versus Engagement

In Table 6, we present estimates of the relationship between teachers' classroom practices and students' math test scores and *Engagement and Happiness in Class*. All coefficients are presented as standardized effect sizes. Focusing first on instructional formats and modalities, we find that random assignment to a teacher who spends more time on combined *Small-Group, Partner, or Independent Work* with components of *Teacher-Led Instruction* (e.g., in centers), relative to teacher-led instruction alone, results in improved student-reported *Engagement* (effect size is 0.29 SD). The effect size linking time spent on *Whole-Class Discussion* and *Engagement* is smaller (0.11 SD) but also positive and substantively meaningful. However, we observe an opposite pattern with regard to effects of time spent on these formats on math test scores, particularly for the project math test where estimates on student-oriented formats relative to teacher-led instruction are negative and statistically significant; for state math test scores, the estimate on *Whole-Class Discussion* also is negative, though small (-0.04 SD) and statistically significant at the $p = 0.1$ threshold. In the unconditional models where the relationship between each practice and student outcomes is not conditioned on the other practices (see Appendix Table 3), the signs of estimates generally are the same but the magnitudes often attenuate towards zero. This pattern occurs because instructional formats are correlated with other practices that also predict student outcomes, as we describe next. Relative to estimates from the conditional models, standard errors are larger in the unconditional models because there is more unexplained variation in the outcome.

We observe similar tradeoffs with regard to math-specific teaching practices and effects on test scores versus student engagement. Random assignment to a teacher scoring 1 SD above the mean on *Ambitious Mathematics Instruction* produces substantive and statistically significant gains in students' math test scores between 0.11 and 0.17 SD, but decreases in *Engagement* of 0.37 SD. Similar to above, this latter estimate attenuates substantially in the model that does not condition on the other teaching practices and is no longer statistically significant, reflecting the fact that *Ambitious Mathematics Instruction* is positively correlated with *Whole-Class Discussion* which then is positively related to student *Engagement*. That said, the unconditional estimate still is negative in magnitude (-0.14 SD). Further, we find that having a teacher who makes errors in their presentation of content (i.e., higher scores on *Mathematical*

Errors) results in a decline in math test scores between 0.09 and 0.14 SD, but an increase in student *Engagement* of 0.29 SD. Here too, the latter estimate attenuates in models that do not condition on other practices, though remains statistically significant (0.23 SD). We believe it is unlikely that students recognized teachers' errors and became engaged because of that. Instead, we infer that mathematical errors may be correlated with other practices. For example, we observe that *Mathematical Errors* occur more often from teachers who spent more time on *Teacher-Led Instruction* (see Table 3), which is less engaging for students. Indeed, when we estimate the relationship between *Mathematical Errors* and student *Engagement* conditioning only on instructional formats, the estimate attenuates closer to zero (0.15 SD). Similarly, there may be other practices not captured on the MQI and CLASS that correlate with *Mathematical Errors* and drive lower student *Engagement*.

For general teaching practices, findings do not reveal the same tradeoffs with regard to improving one student outcome versus another, though patterns are not necessarily consistent. Random assignment to a teacher with strong *Classroom Organization* skills produces positive gains in state and project math test scores between 0.1 and 0.13 SD, but no effect on *Engagement*. We also find a positive effect on student *Engagement* of random assignment to a teacher 1 SD above the mean on *Emotional Support* (0.1 SD), but only in the conditional model. In unconditional models, we also observe a positive relationship between the *Student Engagement* item from the CLASS instrument and student-reported *Engagement* (0.12 SD). However, counter to theory and intuition, we observe a statistically significant and negative relationship between teachers' *Emotional Support* and students' math test scores in the conditional models (-0.08 to -0.09 SD). These estimates attenuate to zero when not conditioning on other teaching practices, likely capturing collinearity between *Emotional Support* and *Classroom Organization* (see Table 3). Finally, we do not find a relationship between test preparation and student outcomes in any model, though the directions of the estimates are expected (positive for the state math test, negative for *Engagement*).

Unpacking Tradeoffs through Qualitative Observations of Instruction

Our qualitative analyses and findings aim to provide insight into several tradeoffs and counterintuitive relationships documented in the quantitative analyses between classroom practices that improve students' math test scores versus classroom engagement. Ultimately, in pairing the quantitative and qualitative analyses, we aim to uncover a set of practices that may drive improvements in both student outcomes.

In total, our observations of teachers and lessons identified 65 unique codes, which we organized into 24 parent codes (see Appendix Table 4). Of the parent codes, we identified 20 as themes (see Table 7) because they captured elements of instruction for multiple lessons for a given teacher and for multiple teachers. Many but not all of these themes align with dimensions of instructional practice described in the MQI and CLASS instruments. We organize the parent codes and resulting themes into three broad categories of instructional activities described by Pianta and Hamre (2009)—two of the developers of the CLASS instrument—in their conceptualizing of "good" or high-quality teaching: instructional supports, emotional supports, and organizational supports.

We are interested in those practices that differentiate high- versus low-engagement teachers. Therefore, in Table 7, we tally the number of high- versus low-engagement teachers (n = 6 in each group) for whom each theme was observed. (We remind readers that all 12 teachers excelled at improving students' math performance.) In doing so, we find that five of the 20 total themes differentiate teachers with regard to impacts on students' *Engagement and Happiness in Class* (cells highlighted in gray). Below, we define and provide illustrative examples of these five themes. In the spirit of our sequential explanatory mixed-methods design, we also link these patterns to those from the quantitative analyses. We order our discussion around practices that emerge more versus less frequently in the lessons of high-engagement teachers, and around practices that are more versus less consistent with theory.

Active mathematics. We find that many of the teachers in our qualitative sample whose students self-reported high scores on *Engagement and Happiness in Class* (*n* = 5 of 6) employed *Active Mathematics* across their lessons, while fewer (*n* = 3 of 6) of the low-engagement teachers engaged in similar activities. We define *Active Mathematics* as lessons in which students engaged in mathematical activities that encouraged hands-on participation, physical movement, or peer interaction. In these lessons, students frequently worked collaboratively in groups, pairs, or math centers/stations. Additionally, these lessons often featured tactile components such as use of manipulatives or engagement in math games (e.g., using die, spinners, etc.). For example, in one lesson, students used egg cartons and counters to find equivalent fractions. The prevalence of *Active Mathematics* in these classrooms is in-line with our hypothesis that activities that promote students' collaboration around hands-on activities are likely to promote engagement in class activities.

Active Mathematics generally occurred in small-group, partner, or independent work, and thus overlaps with this instructional format code from the MQI that was positively related to student engagement. At the same time, two other qualitative codes, *Student-Centered Instruction* and *Teacher-Centered Instruction*, did not differentiate between high- and low-engagement teachers, suggesting that the specific activities that occur in small groups likely matter more than the formats alone. This pattern is consistent with our quantitative analyses that found that the relationship between instructional formats and student outcomes differed when conditioning/not conditioning on other practices.

Similarly, we compare our *Active Mathematics* code to the *Ambitious Mathematics Instruction* dimension from the MQI, where these is some but not complete overlap. For example, when teachers link between multiple mathematical representations to build conceptual understanding (one item from *Ambitious Mathematics Instruction*; see Appendix Table 2), this often includes use of manipulatives. Comparatively, solving the same mathematical problem in two ways and drawing explicit links between these solution strategies would be considered an ambitious, inquiry-oriented activity but would not necessarily require student movement. These distinctions may explain why *Active Mathematics* tended to show up in high-engagement but not low-engagement teachers' lessons in our qualitative analyses, while our quantitative analyses showed negative relationships between *Ambitious Mathematics Instruction* and students' *Engagement* in some models. Further supporting this hypothesis, our qualitative analyses identified another code, *Conceptually-Oriented Instruction*, that identified instances where students were provided opportunities to make sense of the underlying meaning of the mathematics. This code is more similar to *Ambitious Mathematics Instruction* than *Active Mathematics*, and did not differentiate high- versus low-engagement teachers; nor did this code emerge for many teachers (n = 3 in total).

Established practices and routines. A second theme that differentiated high- versus low-engagement teachers was use of *Established Practices and Routines* to organize the classroom and students. Most high-engagement teachers (n = 4 of 6) demonstrated *Established Routines and Procedures* across their lessons, while few (n = 2 of 6) low-engagement teachers did so. We observed that high-engagement teachers drew on classroom management strategies and procedures that promoted efficiency and order in the classroom. For example, teachers often communicated their expectations for student behavior at the beginning of the lesson. This proactive rather than reactive approach seemed to reduce the amount of time actively spent on behavior issues during the lesson. Moreover, the time that teachers did spend on student behavior typically involved short redirections that did not interrupt the flow of the lesson. Another aspect of this theme was teachers' intentional regulation of the pace of the lesson. To do this, teachers often employed timers or other timing strategies. For example, one teacher used music to measure time in the number of songs. While the pace among lessons varied, the teachers seemed intentional about the amount of time spent on activities.

In our quantitative analyses, we found that a closely related construct, *Classroom Organization*, resulted in higher math test scores but was not related to student *Engagement*. Differences in patterns likely are explained by the fact that we conditioned our qualitative sample on teachers' ability to improve students' math performance. In other words, it may be that established routines and organizational techniques result in improved engagement when teachers also are successful at improving math test scores.

Mathematical errors or imprecisions. The presence of *Mathematical Errors or Imprecisions* was another theme that differentiated the lessons of high- versus low-engagement teachers. This code applied to only a few of the high-engagement teachers (n = 2 of 6) and to a majority of the low-engagement teachers (n = 4 of 6). During lessons coded for *Mathematical Errors or Imprecisions*, teachers often made one or more errors. Most often these errors were not egregious, and teachers would sometimes notice and correct them immediately. For example, in one lesson a teacher first referred to a pictured triangle as scalene and then corrected to accurately say it was obtuse. In other lessons, there was sloppiness or lack of clarity in the presentation on the content. One teacher incorrectly identified a rhombus as a regular polygon, and another neglected to refer to fraction parts as equal.

This finding contrasts with our quantitative findings, which found that random assignment to a teacher who made more mathematical errors and imprecision—as scored on the MQI—resulted in higher classroom engagement. One explanation for this difference in patterns may relate to the restriction of our qualitative sample to teachers who were successful at raising math test scores: it may be conditional on teachers' ability to improve test scores that more errors leads to lower engagement. Another explanation may stem from a distinction between our qualitative code and the errors dimension from the MQI. The latter does not count as a mathematical error or imprecision instances in which teachers correct it, while our qualitative code counted these as errors. The reason for this is that we originally identified a second qualitative code, *Teacher Acknowledges and Normalizes Mistakes*, in instances in which the teacher noted and corrected mistakes they or students made in a way that normalized mistakes and often used them as teaching opportunities. This code applied to two teachers (one low-engagement and one high-engagement), but only was observed in one lesson each; therefore, we do not identify it as a theme.

Standards and assessment. A fourth theme that differentiated high- versus low-engagement teachers was the incorporation of content that focused on or referred to *Standards and Assessments*. In total, six of the twelve teachers evidenced this theme across their lessons. This code applied to most teachers (n = 4 of 6) from the bottom tercile of student-reported *Engagement*. In lessons coded with *Standards and Assessments*, teachers often referenced an upcoming standardized test to emphasize the importance of a topic, and in some cases incorporated test formats (e.g., multiple-choice items) into lesson activities. In other lessons, teachers began with a warm-up that reviewed concepts that would appear on an upcoming test, and in some cases these warm-up activities were disconnected from the rest of the lesson. We initially hypothesized that teachers who focused on standardized tests and state standards could be successful in raising test scores by engaging in test-preparation instruction that inspires little intrinsic motivation and may even cause student anxiety. Our qualitative analysis provides some evidence that this could be the case. Relationships generated from our quantitative analyses point in this direction, but the estimates are not statistically significant.

Supportive relationships. The fifth and final theme that emerged from our qualitative analyses and differentiated high- and low-engagement teachers was *Supportive Relationships*. Surprisingly, though, the majority of teachers (n = 4 of 6) for whom this code applied across lessons were in the bottom tercile of student *Engagement and Happiness in Class* scores. In these lessons, teachers would often prompt students to help and support one another. For example, in one lesson a teacher required that early finishers check in with their peers to "encourage and help one another." In other instances, teachers would encourage students to do their best and reassures them that "they can do it." While these results seem counterintuitive, we remind readers that in our quantitative analyses, we found that random assignment to a teacher who scored higher on the *Emotional Support* dimension from the CLASS resulted in decreased math test scores, at least when conditioning on other teaching practices.

To explore a possible explanation for this counterintuitive pattern, we looked for overlap between this code and others. We find that the four low-engagement teachers who were coded for *Supportive Relationships* also were coded for *Standards and Assessment*. (We do not observe the same overlap for any other code.) In the quantitative analyses, we also observe a positive correlation between *Emotional Support* and teacher-reported engagement in test-preparation. It may be that *Supportive Relationships* is associated with other features of classroom practice that we our protocol did not pick up on, and that the other features drive student outcomes.

Discussion

Like many other scholars of teaching and teacher education (e.g., Ball & Forzani, 2009; Grossman et al., 2009; McDonald et al, 2013; Pianta & Hamre, 2009), we began our study under the premise that students' test-based achievement in math—particularly on tests that assess higher-level thinking—and their engagement in math classes and activities are valuable and worth teaching towards. Also aligned to the prior literature (e.g., Lampert, 2011; Cohen, 2011; Kraft, 2019), we document how the pursuit of multiple educational outcomes is challenging and complex. Teachers who are effective at improving students' math achievement, on average, are much less successful at engaging students in class (and vice versa). We also find that some of the teaching practices that improve math test scores (e.g. conceptually based instruction) decrease student-reported engagement. At the same time, our unique data set afford valuable insights into classroom mechanisms that counter this trend. Linking the quantitative analyses with fine-grained qualitative observations of classroom instruction reveal a set of "good" teaching practices that benefit both student outcomes, which in turn can inform practice-based teacher education and professional learning.

Across quantitative and qualitative analyses, there is indication that strong classroom organization and clear routines and procedures may be one place to start. In our quantitative work, random assignment to a teacher who earned high scores on the *Classroom Organization* dimension from the CLASS instrument resulted in improved math performance, on both the state and project assessments. Though we did not find any quantitative link between *Classroom Organization* and student-reported *Engagement*, the qualitative analyses indicate that teachers who excelled both at improving math achievement and engaging students in the classroom often exhibited strong routines and procedures. Some may view these qualitative and quantitative patterns as contradictory. However, we remind readers that we designed the qualitative analyses purposefully to help us understand tradeoffs that emerged from the quantitative results. By focusing qualitative observation on a subset of teachers who excelled at raising math test scores but varied in terms of impacts on student engagement, we set ourselves up to identify teaching practices that simultaneously support both sets of student outcomes.

A key feature of our findings related to classroom organization is that the techniques we observed were proactive rather than reactive, allowing teachers to maintain order and ensure efficient use of time without protracted disruptions. The proactive nature of the routines also meant that teachers maintained order without creating a negative classroom climate by, for example, calling out individual students in an exclusionary way. We reiterate this point, as we recognize the concerns raised by scholars—and that we hold ourselves—around "no-excuses" instructional models. By providing step-by-step procedures to enact in classrooms and in response to student misbehaviors (e.g., Lemov, 2010), no-excuses models aim to address concerns that disruptive behaviors interfere with teachers' instruction. But, these models have started to fall out of favor in recent years (Prothero, 2019), due in part to the exclusionary nature of these approaches that disproportionately affect students with disabilities and students of color (Milner & Tenore, 2010), as well as growing consensus that "good" teaching cannot be reduced to "simple selection of specific moves" (McDonald et al., 2013, p. 380).

What we observed in classrooms was different: teachers appeared quite thoughtful and sophisticated in their use of routines to maintain efficiency and order across the classroom.

Another tension that our findings help unpack relates to conceptually oriented instruction. For decades, teacher educators and teacher education scholars—particularly in mathematics—have prioritized conceptually oriented practices that support students' sensemaking activities (NCTM, 2014; NGA, 2010). In support of this vision, our quantitative analyses show statistically significant and substantively meaningful effects of random assignment to a teacher who earns high scores on *Ambitious Mathematics Instruction* and students' math test scores. However, we also observe negative relationships between this measure and student-reported *Engagement and Happiness in Class*, which raises renewed questions about placing ambitious teaching at the forefront of reform (e.g., Lampert, 2001; Kennedy, 2005; Kirschner et al., 2006). If ambitious practices support some but not all desirable student outcomes, then they may require some reconceptualization.

Our qualitative analyses provide direction. We observe benefits of *Active Mathematics* activities, which overlaps with but is not the same as *Ambitious Mathematics Instruction*. Based on our observations of classrooms, active mathematics emphasizes tactile learning and work in small groups or math centers. Ambitious teaching often includes tactile use of representations and student-to-student interactions, as long as they are in the direct service of building conceptual understanding. Thus, it may be that to promote both students' test-score achievement and engagement, teachers and teacher educators may focus on the active component of ambitious teaching specifically. For example, teachers could support students' conceptual understanding and mathematical sensemaking via curated manipulatives and student-to-student activities. In further support of this claim, qualitative codes on *Student-Centered Instruction* (without qualifying what students were working on) and *Conceptually-Oriented Mathematics* (without qualifying the format where this occurred) did not differentiate high- versus low-engagement teachers. Relatedly, in our quantitative analyses, we found evidence that student-oriented work (often in centers) predicted improved student engagement; while this measures also predicted worse project math test scores, the relationship differed depending on whether other practices—including ambitious teaching—were included in the model. Together, we interpret these patterns as evidence that the active component of ambitious teaching may matter most for supporting both student outcomes.

We also turn to a challenge and tension related to teachers' emotional supports for students that is unresolved in our analyses. Consistent with theory, we find evidence that random assignment to a teacher who earned higher scores on the *Emotional Support* dimension and the *Student Engagement* item from the CLASS benefits student-reported *Engagement and Happiness in Class*. At the same time, there is some evidence in our quantitative analyses that *Emotional Supports* results in decreased math achievement. Further, in our qualitative analyses, the *Supportive Relationships* code emerged more with low- rather than high-engagement teachers. We do not interpret these findings as indication that teachers' interpersonal relationships do not matter. We agree with others' core beliefs on the purpose of education to build strong social functioning, and teachers' interpersonal dynamics as a means of doing so (Pianta & Hamre, 2009). It may be that classrooms with a focus on *Supportive Relationships* have other features that drive student outcomes. This pattern may also stem from our qualitative sampling design, where we selected teachers who all excelled at improving math test scores but varied in terms of their impacts on student engagement. We may have observed different patterns if our qualitative sampling design allowed for variation in terms of teachers' impacts on test scores. Future research may probe this design decision, and patterns related to emotional supports more broadly.

Conclusion

Our study describes teaching both as a multidimensional and a messy endeavor. Attending to the multiple components of student development requires much of teachers' knowledge and practice, and much of teacher educators to support teachers in building these skills. Yet, inside the messiness there is also clarity. "Good" teachers and "good" teaching practices build students' test-based achievement, classroom engagement, and other dimensions. Achieving these goals requires a combination of instructional, emotional, and organizational supports, likely with an emphasis on active classroom activities and proactive routines and procedures.

Finally, teacher education research has historically been grounded in rich description of teachers and teaching, often from classroom observations. This work affords deep understanding of classroom practice but is limited in its ability to generalize on a large scale. On the other hand, experimental designs that identify the effect of teachers on student outcomes often provide limited insight into instructional mechanisms driving these effects. Our study brings together these traditions to reveal and offer insight into a tension surrounding "good" teaching. Ultimately, we advocate for more mixed-method research that affords educators practical guidance to address complex issues in teacher education. This will require that researchers thoughtfully design studies and amass datasets that can both provide causal inference as well as illuminate classroom practice, while thinking creatively about experimental research designs that are possible with and responsive to the realities of the classroom.

References

Ainsworth, M. D. S., Blehar, M. C., Waters, E., & Wall, S. N. (2015). *Patterns of attachment: A psychological study of the strange situation*. Psychology Press.

Ball, D. L., & Cohen, D. K. (1999). Developing practice, developing practitioners: Toward a practice-based theory of professional education. In G. Sykes & L. Darling-Hammond (Eds.), *Teaching as the learning profession: Handbook of policy and practice* (pp. 3-22). Sn Francisco: Jossey-Bass.

Ball, D.L., & Forzani, F. M. (2009). The work of teaching and the challenge for teacher education. *Journal of Teacher Education*, *60*(5), 497-511.

Ball, D. L., Hill, H. C., & Bass, H. (2005). Knowing mathematics for teaching: Who knows mathematics well enough to teach third grade, and how can we decide?. *American Educator,* 14-46.

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.

Betts, J., Appleton, J. J., Reschly, A. L., Christenson, S. L., & Huebner, S. (2010). A study of the reliability and construct validity of the Student Engagement Instrument (SEI) across multiple levels. *School Psychology Quarterly*, *25*(2), 84-93.

Blazar, D. (2015). Effective teaching in elementary mathematics: Identifying classroom practices that support student achievement. *Economics of Education Review*, *48*, 16-29.

Blazar, D. (2018). Validating teacher effects on students' attitudes and behaviors: Evidence from random assignment of teachers to students. *Education Finance and Policy*, *13*(3), 281-309.

Blazar, D., Braslow, D., Charalambous, C. Y., & Hill, H. C. (2017). Attending to general and mathematics-specific dimensions of teaching: Exploring factors across two observation instruments. *Educational Assessment*, *22*(2), 71-94.

Blazar, D., & Kraft, M. A. (2017). Teacher and teaching effects on students' attitudes and behaviors. *Educational Evaluation and Policy Analysis*, *39*(1), 146-170.

Blazar, D., Litke, E., & Barmore, J. (2016). What does it mean to be ranked a "high" or "low" value-added teacher? Observing differences in instructional quality across districts. *American Educational Research Journal*, *53*(2), 324-359.

Bodovski, K., & Farkas, G. (2007). Mathematics growth in early elementary school: The roles of beginning knowledge, student engagement, and instruction. *The Elementary School Journal*, *108*(2), 115-130.

Callejo, M. L., & Zapatera, A. (2017). Prospective primary teachers' noticing of students' understanding of pattern generalization. *Journal of Mathematics Teacher Education, 20*(4), 309-333.

Charalambous, C.Y., & Delaney, S. (2019). Mathematics teaching practices and practice-based pedagogies. In *International Handbook of Mathematics Teacher Education: Volume 1: Knowledge, Beliefs, and Identity in Mathematics Teaching and Teaching Development* (pp. 355-390).

Cobb, P., & Jackson, K. (2011). Towards an empirically grounded theory of action for improving the quality of mathematics teaching at scale. *Mathematics Teacher Education and Development*, *13*(1), 6-33.

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

Cohen, D. K., & Hill, H. C. (2008). *Learning policy: When state education reform works*. Yale University Press.

Connell, J. P., & Wellborn, J. G. (1991). Competence, autonomy, and relatedness: A motivational analysis of self-system processes. In R. Gunnar & L. A. Sroufe (Eds.), *Minnesota symposia on child psychology* (Vol. 23, pp. 43–77). Hillsdale, NJ: Lawrence Erlbaum.

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, *19*(1), 15-18.

Creswell, J. W., Plano Clark, V. L., Gutmann, M. L., & Hanson, W. E. (2003). Advanced mixed methods research designs. *Handbook of mixed methods in social and behavioral research*, *209*, 240.
Emmer, E. T., & Stough, L. M. (2003). Classroom management: A critical part of educational psychology, with implications for teacher education. In *Educational psychologist* (pp. 103-112). Routledge.
Fenstermacher, G. D., & Richardson, V. (2005). On making determinations of quality in teaching. *Teachers College Record*, *107*(1), 186-213.
Forzani, F. M. (2014). Understanding "core practices" and "practice-based" teacher education: Learning from the past. *Journal of Teacher Education*, *65*(4), 357-368.
Grossman, P., & McDonald, M. (2008). Back to the future: Directions for research in teaching and teacher education. *American Educational Research Journal*, *45*(1), 184–205.
Hanushek, E. A. (2002). *The long run importance of school quality* (NBER Working Papers 9071). Cambridge, MA: National Bureau of Economic Research.
Hickman, J. J., Fu, J., & Hill, H. C. (2012). *Creation and dissemination of upper-elementary mathematics assessment modules.* Princeton, NJ: ETS.
Hill, H. C., Blunk, M. L., Charalambous, C. Y., Lewis, J. M., Phelps, G. C., Sleep, L., & Ball, D. L. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and instruction*, *26*(4), 430-511.
Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher*, *41*(2), 56-64.
Ho, A. D., & Kane, T. J. (2013). The Reliability of Classroom Observations by School Personnel. Research Paper. MET Project. *Bill & Melinda Gates Foundation*.
Holmstrom, B., & Milgrom, P. (1991). Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *JL Econ. & Org.*, *7*, 24.
Ivankova, N. V., Creswell, J. W., & Stick, S. L. (2006). Using mixed-methods sequential explanatory design: From theory to practice. *Field Methods*, *18*(1), 3-20.
Jones, Leo. (2007). *The student-centered classroom.* Cambridge University Press.
Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). *Have we identified effective teachers? Validating measures of effective teaching using random assignment*. Seattle: The Bill and Melinda Gates Foundation.
Kennedy, M. M. (2005). *Inside teaching: How classroom life undermines reform*. Harvard University Press.
Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist*, *41*(2), 75-86.
Koedel, C., Mihaly, K., & Rockoff, J. E. (2015). Value-added modeling: A review. *Economics of Education Review*, *47*, 180-195.
Kolluri, S. (Online 2018). Student perspectives on the Common Core: The challenge of college readiness at urban high schools. *Urban Education*.
Kraft, M. A. (2019). Teacher Effects on Complex Cognitive Skills and Social-Emotional Competencies. *Journal of Human Resources.*
Krueger, R. A. (1994). *Focus groups: A practical guide for applied research* (2nd ed.). Thousand Oaks, CA: Sage.
Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied linear statistical models* (Vol. 5). Boston: McGraw-Hill Irwin.
Lampert, M. (2001). *Teaching problems and the problems of teaching*. Yale University Press.
Learning Mathematics for Teaching Project. (2011). Measuring the mathematical quality of instruction. *Journal of Mathematics Teacher Education*, *14*, 25-47.

Lemov, D. (2010). *Teach like a champion: 49 techniques that put students on the path to college (K-12)*. John Wiley & Sons.
Lortie, D. C. (1975). *Schoolteacher: A sociological study*. University of Chicago Press.
Lynch, K., Chin, M., & Blazar, D. (2017). Relationships between observations of elementary mathematics instruction and student achievement: Exploring variability across districts. *American Journal of Education*, *123*(4), 615-646.
McDonald, M., Kazemi, E., & Kavanagh, S. S. (2013). Core practices and pedagogies of teacher education: A call for a common language and collective activity. *Journal of Teacher Education*, *64*(5), 378-386.
Milner IV, H. R., & Tenore, F. B. (2010). Classroom management in diverse classrooms. *Urban Education*, *45*(5), 560-603.
National Council of Teachers of Mathematics (2014). *Principles to actions: Ensuring mathematical success for all.* Reston, VA: Author.
National Governors Association. (2010). *Common Core State Standards.* Author.
National Mathematics Advisory Panel. (2008). *Foundations for success: The final report of the National Mathematics Advisory Panel.* U.S. Department of Education.
Paris, S. G., & Paris, A. H. (2003). Classroom applications of research on self-regulated learning. In *Educational psychologist* (pp. 89-101). Routledge.
Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher*, *38*(2), 109-119.
Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom Assessment Scoring System™: Manual Upper-Elementary*. Paul H Brookes Publishing.
Prothero, A. (2019, March). 'No-excuses' charter schools may be falling out of favor, report suggests. *Education Week*,.
Rittle-Johnson, B., & Alibali, M. W. (1999). Conceptual and procedural knowledge of mathematics: Does one lead to the other?. *Journal of Educational Psychology*, *91*(1), 175.
Sass, T. R. (2008). *The stability of value-added measures of teacher quality and implications for teacher compensation policy*. Brief 4. National Center for Analysis of Longitudinal Data in Education Research.
Spillane, J. P., & Thompson, C. L. (1997). Reconstructing conceptions of local capacity: The local education agency's capacity for ambitious instructional reform. *Educational Evaluation and Policy Analysis*, *19*(2), 185-203.
Steinberg, M. P., & Garrett, R. (2016). Classroom composition and measured teacher performance: What do teacher observation scores really measure?. *Educational Evaluation and Policy Analysis*, *38*(2), 293-317.
Stockard, J., Wood, T. W., Coughlin, C., & Rasplica Khoury, C. (2018). The effectiveness of direct instruction curricula: A meta-analysis of a half century of research. *Review of Educational Research*, *88*(4), 479-507.
Stronge, J. H., Ward, T. J., & Grant, L. W. (2011). What makes good teachers good? A cross-case analysis of the connection between teacher effectiveness and student achievement. *Journal of Teacher Education*, *62*(4), 339-355.
Veenman, M. V., Kok, R., & Blöte, A. W. (2005). The relation between intellectual and metacognitive skills in early adolescence. *Instructional Science*, *33*(3), 193-211.

Tables

Table 1
Demographic Characteristics of Participating Teachers and Students

	Full Project Sample	Experimental Sample	
Female	0.85	0.84	
African-American	0.22	0.19	
Asian	0.03	0.04	
Hispanic	0.03	0.02	
White	0.65	0.67	
Mathematical Content Knowledge (Standardized)	0.01	0.03	
Alternative Certification	0.08	0.06	
Teaching Experience (Years)	10.59	12.29	~
P-value on Joint Test of Significance		0.938	
Teachers	309	53	
Female	0.50	0.49	
African-American	0.41	0.37	*
Asian	0.08	0.12	**
Hispanic	0.24	0.23	
White	0.24	0.25	
FRPL	0.65	0.67	
SPED	0.11	0.05	***
LEP	0.21	0.18	*
Prior Year State Math Test (Standardized)	0.08	0.21	***
Prior Year Project Math Test (Standardized)	-0.01	0.09	**
Prior Year State ELA Test (Standardized)	0.07	0.21	***
P-value on Joint Test of Significance		0.000	
Students	9,141	829	

*** $p<0.001$, ** $p<0.01$, * $p<0.05$, ~ $p<0.1$ on difference between the experimental sample and full project sample.

Table 2
Balance Between Randomly Assigned Teacher Effectiveness in Math and Student Characteristics

	Teacher Effects on State Math Test from Randomly Assigned Teacher
Male	-0.002
	(0.003)
African American	0.004
	(0.009)
Asian	0.019
	(0.014)
Hispanic	0.013
	(0.009)
FRPL	0.001
	(0.005)
SPED	-0.027*
	(0.010)
LEP	-0.003
	(0.009)
Prior Achievement on State Math Test	-0.000
	(0.005)
Prior Achievement on Project Math Test	0.007
	(0.005)
Prior Achievement on State ELA Test	-0.007
	(0.005)
P-Value on Joint Test	0.583
Teachers	53
Students	829

Notes: The regression model includes fixed effects for randomization block. For race/ethnicity groups, the left-out category is White. Robust standard errors clustered at the teacher level in parentheses. *** $p<0.001$, ** $p<0.01$, * $p<0.05$, ~ $p<0.1$

Table 3
Descriptive Statistics for Key Independent and Dependent Variables

	Mean	SD	Reliability	Pairwise Correlations												
				Student Outcomes												
				SMT	PMT	EHC	TL	SG	TSG	WCD	AMI	ME	ES	CO	ST	TPA
Test Scores																
State Math Test (SMT)	0.00	1.00	>0.90	1												
Project Math Test (PMT)	0.00	1.00	>0.82	0.70***	1											
Engagement and Happiness in Class (EHC; 1 = strongly agree to 5 = strongly disagree)	4.10	0.85	0.76	0.15***	0.09***	1										
This math class is a happy place for me to be.	3.98	1.13														
Being in this math class makes me feel sad or angry.	4.38	1.11														
The things we have done in math this year are interesting.	4.04	0.99														
Because of this teacher, I am learning to love math.	4.02	1.19														
I enjoy math class this year.	4.12	1.13														
Observation Scores																
Teacher-Led Instruction (TL; 0 = no and 1 = yes)	0.47	0.17	0.66 ; 0.79				1									
Small Group, Partner, or Independent Work (SG; 0 = no and 1 = yes)	0.33	0.12	0.36 ; 0.79				-0.69***	1								
Teacher-Led + Small Group, Partner, or Independent (TSG; 0 = no and 1 = yes)	0.20	0.12	0.54; 0.79				-0.67***	-0.08	1							
Whole-Class Discussion (WCD; 1 = none to 3 = most/all)	1.05	0.08	0.47 ; 0.92				-0.11~	0.10~	0.05	1						
Ambitious Mathematics Instruction (AMI; 1 = not present to 3 = present and sustained)	1.26	0.11	0.74; 0.74				-0.07	0.03	0.06	0.38***	1					
Mathematical Errors (ME; 1 = not present to 3 = present and sustained)	1.12	0.09	0.56; 0.86				0.21***	-0.12*	-0.16**	-0.04	-0.26***	1				
Emotional Support (ES; 1 = low to 7 = high)	6.40	0.40	0.53				-0.06	0.05	0.03	0.11*	0.23***	0.00	1			
Classroom Organization (CO; 1 = low to 7 = high)	4.02	0.41	0.63				0.10~	-0.02	-0.13*	0.00	0.21***	0.07	0.47***	1		
Student Engagement (SE; 1 = low to 7 = high)	5.20	0.42	0.28				0.07	0.00	-0.09	-0.06	0.12*	-0.01	0.61***	0.57***	1	
Test Preparation Activities (TPA; 1 = never or rarely to 4 = daily)	2.40	0.67	0.80				0.05	0.01	-0.08	-0.06	-0.22***	0.22***	0.12*	0.14*	0.04	1
Use Test Items	2.02	0.84														
Incorporate Item Formats	2.41	0.82														
Teach Test-Taking Strategies	2.39	1.01														
Reallocate Time to Tested Subjects	2.83	0.97														
Focus on Bubble Students	2.36	0.97														

Notes: Statistics are generated from all available NCTE data. With the exception of the teacher observation instruments, reliability indices are internal-consistency reliabilities. For math assessments, reliabilities vary across districts, grades, and years (for state math tests), and across test forms (for the project assessment). For the teacher observation insturments, dimensions from the MQI include two reliability measures: adjusted intra-class correlations (ICCs; listed first), and inter-rater agreement rates (listed second). Dimensions from the CLASS instrument include ICCS only, as only one rater rated each lesson using these constructs. See Appendix Table 2 for a list of items from the MQI and CLASS observation instruments.

*** p<0.001, ** p<0.01, * p<0.05, ~ p<0.1

Table 4
Predictive Power of Engagement and Happiness in Class and Math Test Scores from Upper-Elementary School to High School Outcomes

Student Outcomes in Upper-Elementary School	Student Outcomes in High School					
	State Math Test		Control and Relevance of School Work		Future Goals and Aspirations	
Engagement and Happiness in Class	0.006	0.044***	0.193***	0.188***	0.185***	0.186***
	(0.011)	(0.012)	(0.025)	(0.025)	(0.025)	(0.025)
State Math Test	0.628***		0.080**		0.005	
	(0.014)		(0.030)		(0.029)	
Project Math Test		0.450***		0.056*		0.005
		(0.012)		(0.025)		(0.025)
Students	3,329	3,329	1,784	1,784	1,782	1,782

Notes: Estimates in each column come from separate regression models that control for gender, race/ethnicity, eligibility for free or reduced-price lunch, eligibility for special education services, and limited English proficiency status. Students take state tests once in high school, with the specific grade varying across districts. For the two Student Engagement Instrument constructs, we averaged scores across all available years in high school. Estimates presented as standardized effect sizes. Robust standard errors in parentheses.
*** $p<0.001$, ** $p<0.01$, * $p<0.05$, ~ $p<0.1$

Table 5
Pairwise Correlations Between Experimental Teacher Effects on Students' Math Test Scores versus Engagement and Happiness in Class

Teacher Effects on Students'	State Math Test	Project Math Test	Engagement and Happiness in Class
		Panel A: Full Sample	
State Math Test	1.00		
Project Math Test	0.74***	1.00	
Engagement and Happiness in Class	-0.42**	-0.38**	1.00
		Panel B: Exclude Outliers	
State Math Test	1.00		
Project Math Test	0.61***	1.00	
Engagement and Happiness in Class	-0.43**	-0.24~	1.00

*** $p<0.001$, ** $p<0.01$, * $p<0.05$, ~ $p<0.1$. In Panel A, the sample includes all 53 teachers. In Panel B, pairwise correlations exclude outlier teachers with high influence (i.e., Cook's D greater than $4/(n - k - 1)$, where n is the sample size, k is the number of predictors, and 1 is a degrees of freedom correction). Two outliers are excluded from correlations between teachers' effects on *Engageent and Happiness in Class* versus test scores; and five outliers are excluded from correlation between teachers' effects on the two math test scores.

Table 6
Relationship Between Teaching Practices and Student Outcomes

	State Math Test	Project Math Test	Engagement and Happiness in Class
Small Group, Partner, or Independent Work (Left-Out Category = Teacher-Led Instruction Only)	0.013	-0.117*	0.052
	(0.038)	(0.054)	(0.089)
Teacher-Led + Small Group, Partner, or Independent Work	-0.014	-0.177***	0.294***
	(0.027)	(0.041)	(0.076)
Whole-Class Discussion	-0.041~	-0.048	0.105~
	(0.021)	(0.038)	(0.064)
Ambitious Mathematics Instruction	0.178***	0.257***	-0.367***
	(0.040)	(0.066)	(0.086)
Mathematical Errors	-0.102*	-0.072	0.287*
	(0.042)	(0.061)	(0.111)
Emotional Support	-0.091***	-0.058*	0.112~
	(0.023)	(0.024)	(0.066)
Classroom Organization	0.145***	0.113*	0.018
	(0.037)	(0.044)	(0.070)
Test Preparation	0.032	0.036	-0.028
	(0.028)	(0.026)	(0.073)
Teachers	53	53	53
Students	666	666	666

Notes: Estimates in each column come from separate regression models. All models control for student characteristics listed in Table 1, class characteristics from randomly assigned rosters, and fixed effects for randomization block. Estimates arepresented as standardized effect sizes, with robust standard errors clustered at the teacher level in parentheses.

*** p<0.001, ** p<0.01, * p<0.05, ~ p<0.1

Table 7
Instructional Themes from Qualitative Coding

Theme (sorted by category and then alphabetized)	Description	High-Engagement Teachers	Low-Engagement Teachers
Instructional Support			
Active Mathematics	During the lessons, students are engaged in mathematical tasks that encourage students engagement, fun, and movement. Working collaboratively in groups/pairs and tactile components are often features of active mathematics. Tactile components could include: the use of manipulatives and math games, often in "math centers."	5	3
Affirming Student Thinking	The teacher values and validates student thinking around the mathematics via specific praise/comments or by expanding or using student ideas.	2	2
Conceptually-Oriented Mathematics	The instruction and or activities encourage a conceptual understanding of the mathematics. There are opportunities for students to make sense of the underlying meaning of the mathematics.	2	1
Formative Assessment	The teacher monitors students understanding using a variety of techniques. This may include quick and frequent checks for understanding.	1	1
Making Connections	The teacher connects the lesson content to real-world contexts or students prior mathematical knowledge.	3	2
Mathematical Errors and Imprecisions	The teacher makes mathematical errors or imprecisions that s/he may or may not correct. There may be sloppiness around the mathematical language, and inconsistencies in the teacher's expectations for students' mathematical precision.	2	4
Minimizes Difficulty/ Complexity of Mathematics	During his/her instruction, the teacher downplays the difficulty and/or complexity of the mathematics.	1	1
Procedurally-Oriented Instruction	The instruction is characterized by a focus on mathematical procedures and rote practice rather than on conceptual understanding. The teacher tends to pose questions to elicit a right answer rather than promote student understanding or thinking. In general the lesson is characterized by low-cognitive demand instruction and content.	4	4
Standards and Assessments	During the lessons, there is some some focus or reference to assessments or standards. This may be in the form of test prep activities or reference external test makers.	2	4
Student Misconceptions Ignored	During the lessons, the teacher fails to remediate student misconceptions either because the wrong answer is ignored or the teacher simply overrides the wrong answer by providing the right answer her/himself.	2	2
Student-Oriented Instruction	There are opportunities for students to share their ideas with each other and the classroom in general. Teachers will often take up students' ideas in the mathematical discussion. The teacher may encourage "turn-and-talks" or partner shares.	3	3
Teacher-Centered Instruction	The lesson is centered around the teacher's delivery of the mathematical content. The instructional dialogue is dominated by teacher.	3	3
Technology/Multimedia	During the lessons, the teacher purposefully uses components of technology and multimedia to enhance or supplement the mathematics instruction and learning.	1	1
Unclear Expectations	The teacher is unclear in her expectations for students' work. They may provide insufficient directions at the beginning of the lesson or may contradict themselves over the course of the lesson.	1	1
Emotional Support			
Supportive Rellationships	Students support and help one another, often because of teacher's prompting and modeling. There is evidence that the students are comfortable taking risks and making mistakes, and asking for help.	2	4
Inconsistent Affect	This teacher's affect and behavior toward students is inconsistent, and changes in affect towards students are sometimes pronounced. The teacher may use positive, warm affect when talking to some students but harsh, negative affect with others.	2	1
Negative Affect	The teacher's tone and language toward the students are often harsh and characterized by negativity. The teacher may show rudeness toward students.	2	1
Neutral Affect	The teacher's exhibits a neutral tone and affect toward students. The teacher's language is often direct and matter-of-fact.	1	2
Strong Rapport with Students	There is evidence that the teacher connects with students either by situating classwork or drawing on students own contexts. S/he may also connect with students by sharing personal information or anecdotes with students.	3	1
Organizational Support			
Established Practices and Routines	The teacher has established practices and routines to manage the classroom environment and student behavior. This often includes the use of positive reinforcement and strategies to manage the pace of the lesson.	4	2

Note. We highlight in gray themes that differentiate high- veresus low-engagement teachers

Figures

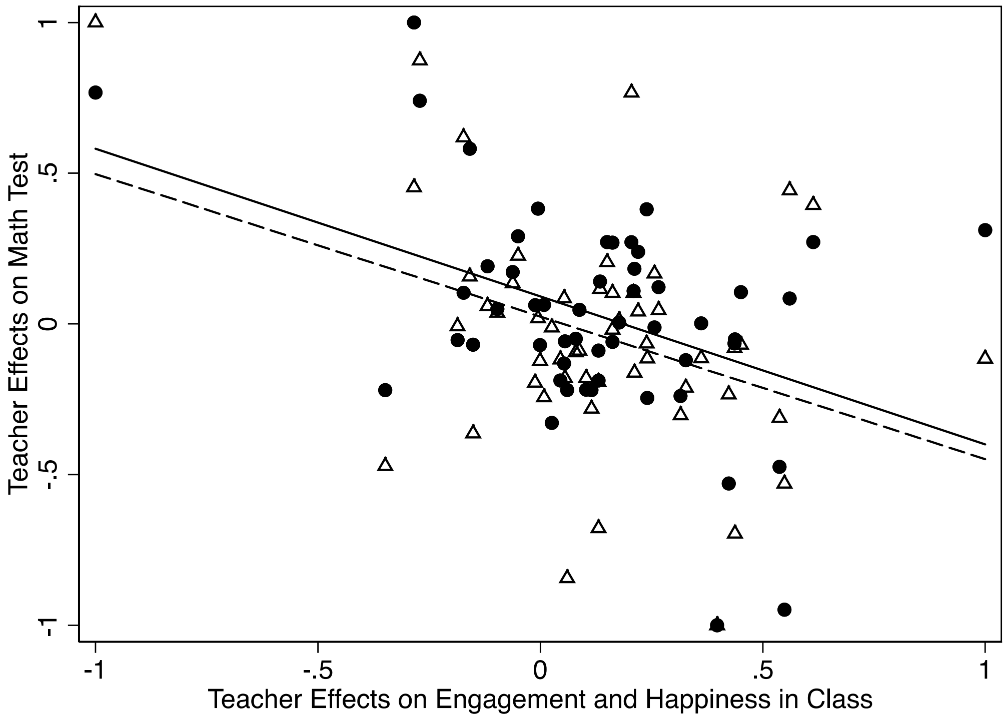

Figure 1. Correlations between Teachers' Effects on Students' Math Performance versus Engagement and Happiness in Class.

Note: Each circle or hollow triangle represents a teacher. Circles and the solid best-fit line represent the relationship between teachers' effects on students' *Engagement and Happiness in Class* and state math test scores; hollow triangles and the dashed best-fit line represent the relationship between teachers' effects on students' *Engagement and Happiness in Class* and project math test scores. Teachers' effects are scaled on range [-1,1] such that the slopes of the best-fit lines represent the correlation coefficients from Table 5.

Appendix

Appendix Table 1
Non-Compliance and Missing Data

Data Source	N	Proportion Missing	*P*-Value on Relationship to Baseline Teacher Effects on State Math Test
	Panel A: Non-Compliance from Full Sample (N = 829)		
Student Non-Compliers	163	0.20	0.222
	Panel B: Missing Data in Compliers Sample (N = 666)		
State Math Test	3	0.00	0.546
Project Math Test	38	0.06	0.960
Engagement and Happiness in Class Survey	22	0.03	0.815
	Panel C: Missing Data on Teachers (N = 53)		
Out-of-Year Observation Scores	2	0.04	0.350
Out-of-Year Test-Preparation Survey	2	0.04	0.350

Appendix Table 2
Item Text for Teacher Observation Instruments

Instrument and Domain	Description
	Mathematical Quality of Instruction
Teacher-Led Instruction	Teacher leads discussion or presentation of mathematical material.
Small Group, Partner, or Independent Work	Teacher divides students into small groups or pairs for work on mathematical problem or task; or students work individually.
Whole-Class Discussion	Teacher is in charge of the class, just as in direction instruction. However, the teacher is not primarily engage in delivering information or quizzing. Rather, she/he has students sharing their thinking, explain their reasoning, and build on one another's contributions.
Ambitious Mathematics Instruction	
Linking and Connections	Linking and connections of mathematical representations, ideas, and procedures.
Explanations	Explanations that give meaning to ideas, procedures, steps, or solution methods.
Multiple Methods	Multiple procedures or solution methods for a single problem.
Generalizations	Developing generalizations based on multiple examples.
Mathematical Language	Mathematical language is dense and precise and is used fluently and consistently.
Remediation	Remediation of student errors and difficulties addressed in a substantive manner.
Use Student Productions	Responding to student mathematical productions in instruction, such as appropriately identifying mathematical insight in specific student questions, comments, or work; building instruction on student ideas or methods.
Student Explanations	Student explanations that give meaning to ideas, procedures, steps, or solution methods.
Student Mathematical Questioning and Reasoning	Student mathematical questioning and reasoning, such as posing mathematically motivated questions, offering mathematical claims or counterclaims.
Enacted Task Cognitive Activation	Task cognitive demand, such as drawing connections among different representations, concepts, or solution methods; identifying and explaining patterns.
Mathematical Errors and Imprecisions	
Major Errors	Major mathematical errors, such as solving problems incorrectly, defining terms incorrectly, forgetting a key condition in a definition, equating two non-identical mathematical terms.
Language Imprecisions	Imprecision in language or notation, with regard to mathematical symbols and technical or general mathematical language.
Lack of Clarity	Lack of clarity in teachers' launching of tasks or presentation of the content.
	Classroom Assessment Scoring System
Emotional Support	
Positive Climate	Positive climate reflects the emotional connection and relationships among teachers and students, and the warmth, respect, and enjoyment communicated by verbal and non-verbal interactions.
Teacher Sensitivity	Teacher sensitivity reflects the teacher's timely responsiveness to the academic, social/emotional, behaioral, and developmental needs of individual students and the entire class.
Respect for Student Perspectives	Regard for student perspectives captures the degree to which the teacher's interactions with students and classroom activities place an emphasis on students' interests and ideas and encourage student responsibility and autonomy. Also considered is the extent to which content is made useful and relevant to the students.
Classroom Organization	
Negative Climate	Negative climate reflects the overall level of negativity among teachers and students in the class.
Behavior Management	Behavior management encompasses the teacher's use of effective methods to encourage desirable behavior and prevent and redirect misbehavior.
Productivity	Productivity considers how well the teacher maages time and routines so that instructional time is maximized. This dimensions captures to degree to which instructional time is effectively managed and down time is minimized for students.
Student Engagement	This scale is intended to capture the degree to which all students in the class are focused and participating in the learning activity presented or facilitated by the teacher. The difference between passive engagement and active engagement is of note in this rating.

Note: Item decsriptions come from the Learning Mathematics for Teaching Project (2011) for the MQI, and from Pianta, La Paro, and Hamre (2008) for the CLASS.

Appendix Table 3
Relationship Between Teaching Practices and Student Outcomes, Not Conditioning on Other Practices

	State Math Test	Project Math Test	Engagement and Happiness in Class
Teacher-Led Instruction	-0.048	0.102	-0.137
	(0.043)	(0.062)	(0.103)
Small Group, Partner, or Independent Work	-0.004	-0.020	-0.017
	(0.027)	(0.045)	(0.048)
Teacher-Led + Small Group, Partner, or Independent Work	0.040	-0.041	0.109
	(0.031)	(0.039)	(0.074)
Whole-Class Discussion	-0.048~	-0.033	0.178***
	(0.026)	(0.043)	(0.035)
Ambitious Mathematics Instruction	0.133**	0.114	-0.137
	(0.047)	(0.082)	(0.104)
Mathematical Errors	-0.143***	-0.114~	0.232*
	(0.037)	(0.062)	(0.087)
Emotional Support	-0.004	-0.005	0.041
	(0.037)	(0.037)	(0.047)
Classroom Organization	0.105**	0.098~	0.005
	(0.034)	(0.058)	(0.081)
Test Preparation	0.043	0.005	0.019
	(0.026)	(0.036)	(0.071)
Student Engagement	-0.023	-0.013	0.123*
	(0.025)	(0.034)	(0.052)
Teachers	53	53	53
Students	666	666	666

Notes: Estimates in each cell come from separate regression models. All models control for student characteristics listed in Table 1, class characteristics from randomly assigned rosters, and fixed effects for randomization block. Estimates arepresented as standardized effect sizes, with robust standard errors clustered at the teacher level in parentheses.

*** $p<0.001$, ** $p<0.01$, * $p<0.05$, ~ $p<0.1$

Appendix Table 4
Codebook for Qualitative Analysis

Parent Code (sorted by category and then alphabetized)	Child Code (alphabetized within parent code)	Description
Instructional Support		
Active Mathematics*		
	Centers and Movement	Students work in centers and/or there are opportunities to move around the room.
	Mathematical Sense-making	The instruction supports students to make mathematical meaning.
	Tactile/Hands-on Activities	The lesson includes hands-on activities or tactile representations of the mathematics (e.g., manipulatives, models, games).
	Variety of Activities	The teacher uses a varieties of activities during the lesson (e.g., manipulatives, white boards, and paper-and-pencil tasks within the same lesson).
Affirming Student Thinking*		
	Direct/Specific Praise	Teacher praises students for specific contributions. Contributions may be mathematical and non-mathematical.
	References and/or Compares Student Ideas	The teacher references student ideas at different times in the lesson or compares ideas from two or more students.
	Revoicing Student Answers	The teacher revoices student answers.
Conceptually-Oriented Mathematics	N/A	The instruction and or activities encourage a conceptual understanding of the mathematics.
Formative Assessment*		
	Frequent Understanding Checks	The teacher informally checks for student understanding (e.g., via hand gestures, exit tickets, etc.)
	Monitoring Work and Participation	The teacher monitors student participation in the classroom, noting work students produce.
Funneling Questions	N/A	The teacher asks questions that encourage one or two word responses or that drive toward the correct answer (e.g., "So did he earn more or less than 20?")
Making Connections*		
	Connections to Background Knowledge	The teacher connects the previous academic experiences.
	Real World Relevance	The teacher incorporates real-world examples into the lesson (e.g., making a connection between measuring ingredients for a recipe and equivalent fractions).
Mathematical Errors and Imprecisions*		
	Inconsistent Emphasis on Precision	The teacher is inconsistent in the degree of mathematical precisions that they expect of students.
	Teacher Errors	The teacher makes a mathematical error or imprecision (e.g., referring to an isosceles triangle as scalene).
Minimizes Difficulty/ Complexity of Mathematics*	N/A	
	Categorizes Problems as Easy or Difficult	The teacher characterizes problems as either easy or difficult.
	Math as Simple/Straightforward	The teacher uses language that reduces the complexities of math to something simple, quick, obvious, etc.
Precise Mathematical Language	N/A	The teacher encourages students' use of precise mathematical language
Procedurally-Oriented Instruction*		
	Emphasis on Practice and Review	The instruction is characterized by low-cognitive demand review and practice.
	Math Tricks	The instruction includes the use of mathematical tricks and mnemonics to help students remember and use math rules (e.g., X comes before Y in an ordered pair just as it does in the
	Preference for Right Answers	The teacher's questions and instruction focus on correct answers, limiting opportunities for student thinking and understanding.
	Rote Practice	The instruction is characterized by a focus on mathematical procedures and rote practice rather than on conceptual understanding.
	Scaffolding Procedures	The instruction focuses on the steps required to execute mathematical procedures.
Standards and Assessments*		
	Connections to Assessments	The teacher notes a connection between the lesson content and an assessment.
	Focus on Standards	The instruction is characterized by a focus on state standards.
	Incentives	The teacher provides incentives/rewards for correct answers (e.g., candy, points, etc.).
	References to Test Materials or Creator	The teacher makes references to test material or material creators (e.g., "They'll want you to show your work.").
	Test Preparation	The teacher incorporates specific test-preparation activities or formats into the lesson activities.
Student Misconceptions Ignored*		
	Corrects and Moves On	The teacher quickly corrects a student error without an explanation of why the error is incorrect.
	Ignores Incorrect Answer	The teacher ignores an incorrect answer, missing an opportunity to address a student misconception.

Student-Oriented Instruction*		
	Active Student Participation	The lesson is characterized by active student participation (e.g., multiple students raising hands, many students engaged in choral responses).
	Grouping Students	Students work in groups.
	Opportunities for Participation	There are opportunities for all students to participate and share their thinking through one or more strategies (e.g., The teacher calls on several students to share their thinking and also has students share their thinking in pairs).
	Productive Struggle	The teacher lets students grapple with mathematical problems.
	Student Choice	The teacher gives students opportunities to make choices during the lesson (e.g., choice in how to solve a problem, choice in partners, etc.)
	Students Offer Original Ideas	Students offer original ideas to explain mathematical concepts.
	Students Pose Questions	Students pose questions to clarify or extend their own learning (e.g., ask teacher to clarify and provide other examples).
	Student-to-Student Talk	Students respond to each others' ideas.
	Turn and Talk	The teacher pairs students to discuss mathematical problems or ideas.
Teacher-Centered Instruction*		
	Limited Student Discussion	The teacher does most of the talking during the lesson, and the teacher's questions do not encourage student discussion.
	Teacher Controlled	The teacher carefully controls the pace and content of the lesson (e.g., early finishers must wait for all students before moving on).
	Teacher Models own Thinking	The teacher thinks aloud as an instructional strategy.
Technology/Multimedia*		
	Multimedia	The teacher uses videos or music in the classroom to support learning.
	Technology Integration	The teacher uses document cameras, slideshows, or some other technology support.
Unclear Expectations*	N/A	The teacher is unclear in expectations for students' work, providing insufficient directions at the beginning of the lesson or contradicting themself over the course of the lesson.
Emotional Support		
Supportive Relationships*		
	Environment of Support	The teacher encourages students to help and support one another.
	Jovial Classroom Culture	The teacher makes jokes and engages the class with humor.
	Personal Connections	The teacher provides information about their personal life.
	Students Solicit Help	Student ask clarifying questions or ask for help.
Inconsistent Affect*	N/A	This teacher's affect and behavior toward students is inconsistent, and changes in affect towards students are sometimes pronounced. The teacher may use positive, warm affect when talking to some students but harsh, negative affect with others.
Negative Affect*		
	Negative/Harsh Language	The affect of the teacher is characterized by negativity and/or harsh language.
	Negative Interactions	There is a negative exchange between the teacher and student(s).
	Negativity Toward Select Students	Teacher shows particularly negative affect/rudeness/harshness toward select students.
Neutral Affect*		
	General Neutral Affect	The teacher's speech and affect with students is generally neutral throughout the lesson.
	Matter-of-Factness	The teacher uses very direct and matter-of-fact language with students.
Positive Affect*	Terms of Endearment	The teacher uses terms of endearment to refer to students.
	Jovial Classroom Culture	The teacher makes jokes and engages the class with humor.
Strong Rapport with Students*		
	Humanization of the Teacher	The teacher shares personal information and/or anecdotes with students (e.g., sharing stories about children, etc.).
	Knowledge of Students	There is evidence that the teacher is knowledgeable about students' preferences and/or academic needs (e.g., "This is your favorite strategy.").
	Quirky Teaching	The teacher engages students into lesson with humor. The teacher makes use of jokes and personal stories to entertain class.
Teacher Acknowledges and Normalizes Mistakes	N/A	The teacher publicly acknowledges and corrects mistakes.
Organizational Support		
Established Practices and Routines*		
	Direct and Behaviorally Focused Classroom Management	The teacher consistently employs classroom management that is direct and behaviorally focused.
	Established Practices	There is evidence of established practices that contribute to the efficiency of the classroom (e.g., students move to different centers efficiently and with little downtime).
	Routines	There is evidence of routines in place to facilitate behavior management.

Note. *Indicates that the parent code is also a theme. Codes qualified as themes if they applied to multiple lessons from multiple teachers/