Journal of Teacher Education
1–16

DOI: 10.1177/00224871231155830
journals.sagepub.com/home/jte
SAGE

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

David Blazar[1] and Cynthia Pollard[2]

Abstract

The pursuit of multiple educational outcomes makes teaching a complex craft subject to potential conflicts and competing commitments. Using a data set in which teachers were randomly assigned to classes paired with videotaped lessons, 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). Furthermore, teaching practices that improve 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, physical movement, and peer interaction; and established routines and procedures to proactively organize the classroom. In addition to guiding practice-based teacher education, our sequential, explanatory 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

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 (Paris & Paris, 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 theoretical and empirical literature on teaching practice—and its inherent dilemmas and predicaments—in three ways. The first contribution is methodological. Like others (Ball & Forzani, 2009; Grossman & McDonald, 2008; McDonald et al., 2013), we argue that building toward a common technical vocabulary requires systematic empirical validation that bridges the strengths of scholars

[1]University of Maryland, College Park, USA
[2]Harvard Graduate School of Education, Cambridge, MA, USA

*Cynthia Pollard is now affiliated to Stanford University, Stanford, CA, USA

Corresponding Author:
David Blazar, Department of Teaching and Learning, Policy and Leadership, University of Maryland, College of Education, 2311 Benjamin Building, College Park, MD 20742, USA.
Email: dblazar@umd.edu

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 classroom practices, and student outcomes (e.g., Hanushek, 2002). We do so by using a data set in which teachers were randomly assigned to class rosters within schools—a rarity in research on teaching, but also a key design strategy for drawing causal claims (Charalambous & Delaney, 2020). Next, we build on the tradition of teacher education researchers to directly observe classroom instruction (e.g., Stronge et al., 2011). Drawing from established observation protocols and open-ended qualitative inquiry, we consider a set of teaching practices that may be responsible for improved student outcomes. Some scholars have identified tensions between these frameworks, noting that "quality" teaching as identified through classroom observation may not be the same practices as those that are "successful" or "effective" at improving student outcomes (Fenstermacher & Richardson, 2005). 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 is empirical theory testing, where the methodology described above provides 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 and ambitious content) can simultaneously result in decreased student engagement.

Our third contribution is to practice: We aim to make sense of the tradeoffs described above to provide direction for the teaching and teacher education fields on a "common technical vocabulary" of practices that can improve both students' test-score performance and classroom engagement. To do so, we combine the quantitative analyses of classroom observation scores with 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 routines and procedures, where teachers communicate their expectations for students in a way that increases efficiency and order. 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). Our findings suggest that when routines are proactive rather than reactive, they can create classroom conditions that benefit students' test scores and engagement.

Other scholars describe hands-on participation and established routines and procedures—and additional classroom practices—as being "high leverage" and central to a "common technical vocabulary."[1] At the same time, the methods and rationale for calling out these practices in our study builds from the prior literature in important ways: We identify these practices because they support multiple student outcomes, and links between these practices and varied student outcomes are identified through a randomized design that also includes a qualitative classroom observation component. Our design also affords an opportunity to partially reconceptualize "good" practices relative to the prior literature, including in particular the active component of "ambitious" teaching versus its focus on conceptual understanding—a topic we turn to in our discussion. Ultimately, we argue that hands-on participation and established routines and procedures 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 article is a long-standing one: What makes for "good" teaching? 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 qualitative examination of classrooms and teacher-student interactions, to quantitatively

linking teachers to student outcomes in large-scale data sets. 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)

Our study builds closely from this perspective, particularly in the push for theory-based empirical analysis. 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 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 repeatedly 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: (a) emotional supports that attend to students' sense of attachment (Ainsworth et al., 1978) and self-determination (Connell & Wellborn, 1991); (b) organizational supports that help students build self-regulatory skills for behavior, time-use, and attention in the classroom (Paris & Paris, 2001); and (c) 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. 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—scholars often have advocated for inquiry-oriented or ambitious practices that elicit and build on students' mathematical thinking to make meaning of mathematical concepts. Professional boards and panels including the National Council of Teachers of Mathematics (2014) and the National Mathematics Advisory Panel (2008) recommend that teachers support students by providing explanations for mathematical phenomena (Leinhardt, 1989), 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 is also associated with calls for student-centered as opposed to teacher-led instruction (Jones, 2007). However, 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 toward multiple, potentially competing goals is further documented in the work of economists seeking to monitor and reward performance. For example, Holmstrom and Milgrom (1991) describe the multitask nature of teachers' work as a potential "problem" because of multiple incentives (e.g., salary, job advancement) that simultaneously advocate for basic skills and strive for higher-level thinking and curiosity. In the face of competing incentives, teachers are pulled in multiple directions. 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 non-random matching of students to teachers. To address this "omitted variables bias" problem, Kraft (2019) used a portion of the Measures of Effective Teaching (MET) project data set in which teachers were randomly assigned to class rosters within schools, finding 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. In similar analyses, Blazar and Kraft (2017) and Blazar (2018) used a subset of the data described in this paper to examine teachers' contributions to students' mathematics test scores versus their contributions to student-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? Aligned 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 are 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? 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 (Kennedy, 2005; Lampert, 2001) 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 (Forzani, 2014; Grossman & McDonald, 2008; 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). A specific illustration of this need relates to long-standing discussion on the benefits of direct versus student-centered instruction. In a recent meta-analysis on this topic that pooled results from all quantitative analyses—no matter the design—the authors concluded the benefits of direct instruction to a range of student outcomes (Stockard et al., 2018). Yet, the much smaller subset of random assignment analyses often returned much smaller or null effects. If the goal of teacher educators is to build toward a common vocabulary and the successful pursuit of multiple student outcomes, then these tensions must be unpacked and resolved.

Ultimately, we argue that, to better understand challenges and tradeoffs of "good" teaching, we must build on the strengths of multiple disciplinary traditions and multiple approaches to research design. Our study fills this gap.

Sample and Experimental Design

Our project leverages a unique data set from the National Center for Teacher Effectiveness (NCTE) collected over a 3-year period (2010–2011 through 2012–2013) that includes fourth- and fifth-grade teachers and their students in four school districts on the east coast of the United States. 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.

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 data set ($n = 309$ teachers 9,141 students) on a host of characteristics collected by the project including education, experience, and content knowledge ($p = .938$ on a joint test of significance comparing the characteristics of teachers in the two groups). Like many districts across the United States, the 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, although 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 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. 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. Of the baseline student characteristics, only students' eligibility for special education

Table 1. Background Characteristics of Participating Teachers and Students.

Background characteristics	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		.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
Free or Reduced-Price Lunch	0.65	0.67
Special eEducation	0.11	0.05***
Limited English Proficiency	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 English Language Arts Test (standardized)	0.07	0.21***
P value on Joint Test of Significance		.000
Students	9,141	829

†$p < .1$. *$p < .05$. **$p < .01$. ***$p < .001$ on difference between the experimental sample and full project sample.

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

Background 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)
Free or Reduced-Price Lunch	0.001
	(0.005)
Special Education	−0.027*
	(0.010)
Limited English Proficiency	−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 English Language Arts Test	−0.007
	(0.005)
P-value on Joint Test	0.583
Teachers	53
Students	829

Note. 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.
†$p < .1$. *$p < .05$. **$p < .01$. ***$p < .001$.

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 ($p = .583$). (For additional details on the sample and experimental design, see Supplemental Appendix A, available with the online version of this article.)

A second possible threat to our ability to causally link teachers and their practices to student outcomes is attrition due to non-compliance among participating students. Of the 829 students in the experimental sample, 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 (for details, see Blazar, 2018). All analyses presented below exclude these students because they no longer were part of primary data collection. Importantly, this restriction should not threaten the internal validity of results because non-compliance is unrelated to the experiment. 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 ($p = .222$; see Supplemental Appendix A Table 1).

Data for Quantitative and Qualitative Inquiry

The data used in this study include a set of student outcomes (i.e., math test scores, student-reported engagement) and teaching practices (i.e., observations of instruction, teacher reports). The student outcome data are used primarily in the quantitative analyses that help us identify challenges and tradeoffs in teachers' work: We examine how teachers contribute to students' test-score outcomes versus student-reported engagement. We also use the student outcome data in our selection of the qualitative sample of teachers who differ in their effectiveness at improving these outcomes. Then, we use the classroom observation data and a teacher survey to help unpack these challenges and tradeoffs: We examine specific classroom practices that support one type of student outcome versus another. We engage in this work quantitatively by using classroom observation scores of the videotaped lessons and teacher-reported practices; we also draw on the

videotaped lessons themselves in more open-ended qualitative analyses.

Student Outcomes

Our data set includes two types of math assessments. One was developed by the NCTE research team in collaboration with the Educational Testing Service, designed to capture students' understanding of upper-elementary mathematics topics: numbers and operations, algebra, geometry, and measurement. Internal consistency reliability (α) is .82 or higher for each test form. 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. The other type of math assessment is state tests, which varied across states and districts in terms of their formats and cognitive demand ($\alpha > .9$ across tests). 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. (For additional details on the test coding analyses, correlations between the different assessments, and missingness, see Supplemental Appendix A).

In addition, 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* (five items; $\alpha = .76$; see Supplemental Appendix A Table 2 for item text and Blazar & Kraft, 2017, for exploratory factor analyses). To create final scores, we averaged response scores across all items and then standardized to a mean of 0 and *SD* of 1. It is possible that students 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. However, in the online appendix, we show that the measure has predictive validity to test scores and self-reported academic expectations in high school (see Supplemental Appendix A Table 3), indicating the measure captures important information about students' schooling trajectories.

Teaching Practices

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 device that recorded both the teacher and students visually and aurally. Each lesson lasted between 45 and 60 min, on average. 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. Each videotaped lesson is accompanied by a transcript.

In addition to examining the videotapes qualitatively, 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 et al., 2008; see Supplemental Appendix A Table 4 for item text). Following instrument developers' protocols, two raters scored each lesson on the MQI and one rater scored each lesson on the CLASS. We generated teacher-level scores, averaging across raters (where applicable) and lessons. (For additional details on scoring protocols, see Supplemental Appendix A.)

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 three formats (i.e., teacher-led, small group, or both) are mutually exclusive and scored 0 = "no" or 1 = "yes." In our sample, 47% of instruction was teacher-led, while 33% was small group, partner, or independent work, and 20% was a combination of both (see Supplemental Appendix A Table 5). Measurement properties are similar to other studies (e.g., Bell et al., 2012) and indicate that the scores adequately capture the constructs of interest: inter-rater agreement [IRA] of instructional modality is 79%, and intra-class correlations [ICC] that capture the percentage of variation at the teacher level as opposed to construct-irrelevant sources of variation are .66, .36, and .54 for the three formats, respectively. *Whole-Class Discussion* is another mode of instruction—not mutually exclusive with the others—capturing instances in which teachers lead the delivery of content 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 (IRA = .92, ICC = .47).

The remaining dimensions are quality rather than quantity codes. Also from the MQI, *Ambitious Mathematics Instruction* 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 (10 items, scored 1 = "not present" to 3 = "present and sustained"; IRA and ICC = .74). *Mathematical Errors* identifies any errors or imprecisions that teachers introduce into the lesson that go uncorrected (three items, with same scale as above; IRA = .86, ICC = .56). Thus, unlike the other dimensions of

practice, higher scores reflect worse instruction. The quality of ambitious teaching is positively correlated with the amount of time teachers spent on discussion ($r = .38$)—both of which were relatively rare in the sample—but not with time spent in small group, partner, or independent work. Errors were more common for teachers who spent more time leading instruction ($r = .21$) and less common for teachers who spent more time having students work in small groups ($r = -.12$) or leading instruction in small groups ($r = -.16$). Teachers who made more errors engaged less in ambitious teaching, capturing in part the fact that the MQI does not allow instructional activities to be coded as ambitious if they are mathematically incorrect ($r = -.26$).

From the CLASS, *Emotional Support* focuses on teachers' interpersonal relationships with students, including the extent to which teachers create a positive classroom climate and teachers' respect for student ideas (3 items, scored 1 = "low" to 7 = "high"; ICC = .53; no IRA because there is only one rater per lesson). *Classroom Organization* captures teachers' behavior management skills and the pacing of the lesson (3 items, with same scale as above; ICC = .63). The CLASS also includes a single item, *Student Engagement*, that instrument developers separate from the other dimensions (ICC = .28, which is expectedly lower than for other constructs given that this is a single item). We include this item in some of our analyses because of its close connection to student-reported *Engagement* outcome measure.[2] Scores from the CLASS are moderately to strongly correlated with each other (r =.47–.61). However, we keep them as separate because the dimensions have distinct theoretical underpinnings (Pianta & Hamre, 2009) and factor analyses on our data identify them as unique (Blazar et al., 2017). In comparison, correlations of CLASS scores with those from the MQI are no higher than 0.23. We standardized all teacher-level observation scores to have a mean of 0 and *SD* of 1.

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"; $\alpha = .80$). 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* ($r = -.22$).

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 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 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 are our teacher effect estimates, which we specify as a set of random effects. Because the randomized design successfully created balanced groups, we can be assured that these estimates capture the contribution of an individual teacher to student outcomes that is not confounded with other factors, beyond those already controlled for in the model. The random-effects model shrinks the teacher effect estimates back toward the mean based on their precision (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. The estimated correlations between teachers' contributions to students' test scores versus engagement are a key contribution of our work, as they provide empirical evidence on the challenges and tradeoffs in teachers' work.

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, which help to unpack the challenges and tradeoffs for teachers in supporting one student outcome versus another. These analyses aim to identify potential mechanisms related to classroom teaching that may explain the tradeoff. 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 can 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 (see Supplemental Appendix A Table 5), 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).

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 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 et al., 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 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 examine 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 were sufficient to identify themes and cross-group differences (Blazar et al., 2016).

Next, we randomly selected three lessons per teacher, guided by a generalizability and decision study conducted by Hill et al. (2012) which identified this number of lessons as appropriate for achieving sufficiently high reliability through observation. 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.[3] 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. After reviewing all lessons for a given teacher, the full research team met as a group to discuss notes, identify features of classroom instruction or environment that were evidenced in multiple lessons of the focal teacher, and write a memo synthesizing the conversation. After repeating this process for all 12 teachers, the authors undertook a data-driven approach to thematic analysis of the memos (Creswell et al., 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 that 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 3, 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, we include all 53 teachers, while in Panel B we exclude bivariate outliers based on Cook's (1977) D. 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.

Consistent with prior work (Blazar, 2018; Blazar & Kraft, 2017), we find that teachers' effects on students' math test scores are negatively correlated with teachers' effects on students' *Engagement and Happiness in Class*. Correlation coefficients range from −.24 to −.43, and 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

Table 3. Pairwise Correlations Between Experimental Teacher Effects on Students' Math Test Scores Versus Engagement and Happiness in Class.

Teacher effects	State math test	Project math test	Engagement and happiness in class
Panel A: Full sample			
State Math Test	1		
Project Math Test	.74***	1	
Engagement and Happiness in Class	−.42**	−.38**	1
Panel B: Exclude outliers			
State Math Test	1		
Project Math Test	.61***	1	
Engagement and Happiness in Class	−.43**	−.24†	1

Note. 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 *Engagement and Happiness in Class* versus test scores; and five outliers are excluded from correlation between teachers' effects on the two math test scores.
†$p < .1$. *$p < .05$. **$p < .01$. ***$p < .001$.

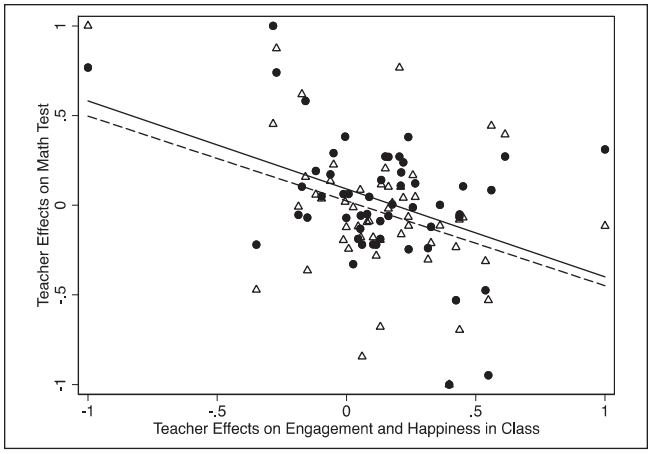

Figure 1. Correlations Between Teachers' Effects on Students' Math Performance Versus Engagement and Happiness in Class.
Note. Each circle or hollow triangle represents a teacher. Circles and the solid best-fit line represent the relationship between teachers' effects on students' *Engagement and Happiness 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].

x- and *y*-axes scaled to [−1,1] so that the slopes of the lines match the correlation coefficients from Panel A of Table 3 that includes the full sample of teachers. 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) capture the set of teachers from which we randomly selected our sample for qualitative analysis.

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). Furthermore, 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 4, we present estimates of the relationship between teachers' classroom practices and students' math test scores versus *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 *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* (0.29 *SD*). The effect size linking time spent on *Whole-Class Discussion* to *Engagement* is smaller (0.11 *SD*) but also positive and substantively meaningful.

However, we observe an opposite pattern with regard to effects on math test scores of time spent on these formats. For effects on the project math test, estimates for student-oriented formats relative to teacher-led instruction are negative and statistically significant. For state math test scores, the estimate for *Whole-Class Discussion* also is negative, though small (−0.04 *SD*) and statistically significant at the $p = .1$ threshold. In the unconditional models where the relationship between each instructional format and student outcomes is not conditioned on the other practices, the signs of estimates generally are the same but the magnitudes often attenuate toward zero (see Supplemental Appendix B Table 1; Supplemental Appendix B, available with the online version of this article, includes all supplemental results). This pattern occurs because instructional formats are correlated

Table 4. Relationship Between Teaching Practices and Student Outcomes.

Teaching practices	State math test	Project math test	Engagement and happiness in class
Small group, partner, or independent work	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

Note. 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 are presented as standardized effect sizes, with robust standard errors clustered at the teacher level in parentheses. For *Small group, partner, or independent work and Teacher-led + small group, partner, or independent work,* the reference category is *Teacher-led instruction only*.
†$p < .1$. *$p < .05$. **$p < .01$. ***$p < .001$.

with other practices that also predict student outcomes, as we describe next.

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*. 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* (see Supplemental Appendix B Table 1). That said, the unconditional estimate still is negative in magnitude (−0.14 *SD*). Furthermore, 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 (−0.09 to −0.14 *SD*), but an increase in student *Engagement* (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 Supplemental Appendix A Table 5), 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. 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, we find that 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*. Random assignment to a teacher with strong *Emotional Support* improves student *Engagement* (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 Supplemental Appendix A Table 5). 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 who all improved math test scores but varied in the extent to which they improved classroom engagement identified 65 unique codes, which we organized into 24 parent codes (see Supplemental Appendix B Table 2). Of the parent codes, we identified 20 as themes (see Supplemental Appendix B Table 3) 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, we tallied the number of high- versus low-engagement teachers ($n = 6$ in each group) for whom each theme was observed (see Supplemental Appendix B Table 3). 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*. Two of these themes—*Active Mathematics* and *Established Practices and Routines*—were observed more in the lessons of high-engagement teachers. The other three themes— *Mathematical Errors and Imprecisions*, *Standards and Assessments*, and *Supportive Relationships*—showed up more in the lessons of low-engagement teachers. Therefore, we begin our discussion with the first two themes, which we argue provide the most direction for practice. In the spirit of our sequential explanatory mixed-methods design, we also link the qualitative themes to patterns from the quantitative results.

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 the following elements: (a) hands-on participation, (b) physical movement, or (c) peer interaction.

Hands-on-Participation. These lessons often featured tactile components such as the use of manipulatives or engagement in math games (e.g., using die, spinners). For example, the lessons from one teacher involved fraction bars and paper folding. Another teacher had students use egg cartons and counters to find equivalent fractions. These activities seemed to support students in bridging concrete and abstract reasoning in support of math performance while also being inherently hands-on for students.

Physical Movement/Student Choice. *Active Mathematics* lessons often included partner work and opportunities to move about the classroom, breaking up instructional time and allowing students to engage their bodies. These lessons also contained many elements of choice (e.g., choice of partner, choice of problem to work on).

Peer Interaction. Lessons featuring *Active Mathematics* often included small group or partner work, allowing students to engage with each other. Sometimes lessons included substantial time on independent work, but allowed students alternative opportunities to engage with their peers (e.g., via student presentations, discussions of board work)

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. At the same time, another qualitative code related to *Active Mathematics—Student-Centered Instruction*—did not differentiate 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 Supplemental Appendix A Table 4), 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 out of 12; see Supplemental Appendix B Table 3).

Established Routines and Procedures. A second theme that differentiated high- versus low-engagement teachers was the use of *Established Routines and Procedures* 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 lessons with *Established Routines and Procedures* were characterized by two elements: (a) proactive classroom management strategies and (b) intentional lesson pacing.

Proactive Classroom Management. To promote order in the classroom, teachers often communicated their expectations for student work and behavior at the beginning of the lesson. For example, teachers would describe the structure of the lesson and what students were expected to do. Teachers also referred back to routines and procedures established earlier. This proactive rather than reactive approach seemed to reduce the amount of time actively spent on redirecting off-task behavior during the lesson, and in general, students were on task and focused throughout class time. For example, one teacher seemed to engage particular students in lines of questioning around the instructional content to forestall off-task behavior. Moreover, the time that teachers did spend on student behavior typically involved short redirections or reminders that did not interrupt the flow of the lesson. The proactive rather than reactive approach also meant that teachers were not calling out individual students in an exclusionary way that can be harmful to that student and to the overall classroom environment.

Intentional Lesson Pacing. Another aspect of this theme was teachers' intentional regulation of the lesson pace, which often was informed by frequent check-ins with students (e.g., non-verbal signals for more time, early finishers). Teachers also often employed timers or music to regulate the pace, for example, using the length of a song to time transitions between activities. While the pace varied among lessons, the teachers seemed intentional about the amount of time spent on activities.

Benefits of *Established Routines and Procedures* that emerge in our qualitative work connect directly to patterns from our quantitative analyses. There, we found that a closely related construct, *Classroom Organization*, resulted in higher math test scores. In the quantitative work, *Classroom Organization* was not related to student *Engagement*. This difference may be related to our research design. Because we conditioned our qualitative sample on teachers' ability to improve students' math performance, 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 lead 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 12 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 general 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 effects on students' *Engagement and Happiness in Class*. 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. In other instances, teachers would encourage students to do their best and reassure them that "they can do it." While these results seem counterintuitive, we observed similar patterns in our quantitative analyses. There, 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 (see Supplemental Appendix A Table 5). It may be that *Supportive Relationships* is associated with other features of classroom practice that 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 & McDonald, 2008; McDonald et al., 2013; Pianta & Hamre, 2009), we began our study under the premise that students' test-based achievement in math and their engagement in math classes and activities are valuable and worth teaching toward. In line with the prior literature (e.g., Cohen, 2011; Kraft, 2019; Lampert, 2001), we document how the pursuit of these multiple educational outcomes is challenging and complex, and contribute new methodological, empirical, and practical insights.

Methodologically, this work illustrates the affordances of applying a sequential, explanatory mixed-method design to the study of teaching practice. Teacher education research has historically been grounded in rich description of teachers and teaching, often from classroom observations. This tradition affords deep understanding of classroom and teaching practice but is limited in its ability to generalize on a large scale. At the same time, experimental designs that identify the effect of teachers on student outcomes often provide limited guidance into instructional mechanisms driving these effects. Our study brings together these traditions to reveal and offer insight into a tension surrounding "good" teaching. The quantitative phase first provided a general understanding of the relationship between student engagement and test scores, while the subsequent qualitative phase helped make sense of that relationship. Ultimately, we advocate for more mixed-method research that affords educators practical guidance to address complex issues in teaching and teacher education. This will require that researchers thoughtfully design studies and amass data sets that can both provide causal inference and illuminate classroom practice.

Empirically, we contribute additional evidence to the collection of studies (e.g., Cohen, 2011; Kraft, 2019; Lampert, 2001) that describe challenges and tradeoffs in the nature of teachers' work. Here, we document how 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.

We delve more in-depth into the empirical work—using the methodological tools described above—that ultimately allows us to speak to practice. Across quantitative and qualitative analyses, there is indication that strong classroom organization and clear routines and procedures may be one place to start in efforts to build a "common technical vocabulary" aimed at improving multiple student outcomes. In our quantitative work, random assignment to a teacher who earned high scores on *Classroom Organization* 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 at 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 procedures, routines, and 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 concerns raised by scholars and practitioners—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, 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 area of classroom practice 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 (National Council of Teachers of Mathematics, 2014; National Mathematics Advisory Panel, 2008). 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*, which raises renewed questions about placing ambitious teaching at the forefront of reform (e.g., Kennedy, 2005; Kirschner et al., 2006; Lampert, 2001). If ambitious practices support some but not all desirable student outcomes, then they may require some reconceptualization.

Pairing these quantitative patterns with qualitative observation provides clarity and direction. We observe benefits of *Active Mathematics* activities, which overlap with but are 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. In further support of this claim, additional qualitative codes on *Student-Centered Instruction*—which does not qualify what students were working on—and *Conceptually-Oriented Mathematics*—which does not qualify 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 measure 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 sets of student outcomes.

Finally, we 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 *Emotional Support* and *Student Engagement* from the CLASS observation instrument benefits student-reported *Engagement and Happiness in Class*. At the same time, there is some evidence in our quantitative analyses that *Emotional Support* results in decreased math achievement. Furthermore, 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 as both 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.

Acknowledgments

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.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

David Blazar https://orcid.org/0000-0001-5596-1552

Notes

1. See, for example, Teaching Works' resource library, which draws on the literature of Ball and Forzani (2009), Grossman & McDonald (2008), and others: https://library.teachingworks.org/curriculum-resources/high-leverage-practices/.
2. As described in the theoretical framing of this article, the Classroom Assessment Scoring System includes another dimension of practice called *Instructional Supports*, which we exclude from our quantitative analyses given theoretical and empirical overlap with the instructional components from the Mathematical Quality of Instruction (Blazar et al., 2017).
3. Due to language in the teacher consent forms, it only was possible for the two authors to watch the videotaped lessons, given our role as researchers on the original project. Additional raters read lesson transcripts, allowed under the consents. Therefore, we made sure that one of the authors and two to three additional raters were assigned to each lesson.

Supplemental Material

Supplemental material for this article is available online.

References

Ainsworth, M. D. S., Blehar, M. C., Waters, E., & Wall, S. N. (1978). *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). 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*, *29*, 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.

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. (2020). Mathematics teaching practices and practice-based pedagogies. In D. Potari & O. Chapman (Eds.), *International handbook of mathematics teacher education. Vol. 1: Knowledge, beliefs, and identity in mathematics teaching and teaching development* (2nd ed., pp. 355–390).

Cohen, D. K. (2011). *Teaching and its predicaments*. 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). 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. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social and behavioral research* (pp. 209–240).

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). National Bureau of Economic Research.

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.

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

Holmstrom, B., & Milgrom, P. (1991). Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics, & Organization*, *7*, 24–52.

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, L. (2007). *The student-centered classroom*. Cambridge University Press.

Kane, T. J., Taylor, E. S., Tyler, J. H., & Wooten, A. L. (2011). Identifying effective classroom practices using student achievement data. *Journal of Human Resources*, *46*(3), 587–613.

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.

Kraft, M. A. (2019). Teacher effects on complex cognitive skills and social-emotional competencies. *Journal of Human Resources*, *54*(1), 1–36.

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.

Leinhardt, G. (1989). Math lessons: A contrast of novice and expert competence. *Journal for Research in Mathematics Education*, *20*(1), 52–75.

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, H. R., IV, & 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*.

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. (2001). Classroom applications of research on self-regulated learning. *Educational Psychologist*, *36*, 89–101.

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.

Prothero, A. (2019, March). "No-excuses" charter schools may be falling out of favor, report suggests. *Education Week*. https://www.edweek.org/policy-politics/no-excuses-charter-schools-may-be-falling-out-of-favor-report-suggests/2019/03

Rittle-Johnson, B., & Star, J. R. (2007). Does comparing solution methods facilitate conceptual and procedural knowledge? An experimental study on learning to solve equations. *Journal of Educational Psychology*, *99*(3), 561.

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.

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.

Author Biographies

Dr. David Blazar is an Associate Professor at the University of Maryland, College Park. His research examines resources that support student outcomes and alleviate inequality, with a particular focus on teachers and teaching quality. He holds an Ed.D., Ed.M., and B.A. from Harvard University, and an M.S.T. from Fordham University. He previously taught high school in New York City.

Dr. Cynthia Pollard is a Postdoctoral Scholar at the Stanford Graduate School of Education. She studies the role that teachers and teaching play in reinforcing or challenging race- and class-based inequality in K-12 classrooms. She holds a Ph.D. and Ed.M. from Harvard University, and a B.A. from the University of California, Los Angeles. She previously taught elementary school in Los Angeles.