INTERNATIONAL LITERACY ASSOCIATION READING RESEARCH QUARTERLY

Effects of Online Content-Focused Coaching on Discussion Quality and Reading Achievement: Building Theory for How Coaching Develops Teachers' Adaptive Expertise

Richard Correnti
Lindsay Clare Matsumura
Marguerite Walsh
Dena Zook-Howell
Donna DiPrima Bickel
Baeksan Yu
University of Pittsburgh, Pennsylvania, USA

The authors conducted a small-scale randomized control trial (n = 31 teachers) of Online Content-Focused Coaching, an intervention consisting of an online workshop followed by multiple cycles of remote video-based coaching, to support dialogic text discussions. Findings demonstrate the efficacy of Online Content-Focused Coaching in three different ways. First, the authors' analyses, after accounting for differential attrition among groups, demonstrate an existence proof for effects of the intervention on both classroom text discussion quality and student achievement. Second, the authors examined and demonstrated an association between the magnitude of changes in discussion quality and the magnitude of achievement gains. Finally, the authors propose and examine evidence to support a theory for how teachers develop adaptive expertise for facilitating dialogic text discussions. Results show that teachers' use of transitional and some aspirational discussion moves grew from baseline to the end of the workshop, with limited growth in the quality of students' contributions. Over the coaching phase of the intervention, teachers' facilitation moves grew substantively, and so did students' strong contributions. The authors interpret the results to suggest that the workshop was critical for developing teachers' knowledge of the features of dialogism and that coach-guided reflection was essential for developing teachers' expertise at using facilitation moves to elicit student thinking. Findings contribute to a validity argument for the efficacy of Online Content-Focused Coaching. More importantly, investigating and describing the process of teaching change is the study's main theoretical contribution.

Classroom discussions that provide opportunities for students to analyze complex texts, actively construct meaning with their peers, and make inferences based on text evidence are a critical component of effective reading comprehension instruction (see Murphy, Wilkinson, Soter, Hennessey, & Alexander, 2009; Wilkinson, Murphy, & Binici, 2015). Such discussions (referred to here as dialogic text discussions) build students' comprehension skills through interactive dialogue and joint sense making around text events and ideas (R. Alexander, 2006). When teachers prompt students to make sense of complex or confusing portions of a text, these social processes ultimately emerge as mental functions that enable students to construct coherent representations of text meaning. Teachers nurture this developmental process by

INTERNATIONAL LITERACY ASSOCIATION

Reading Research Quarterly, 0(0)
pp. 1–40 | doi:10.1002/rrq.317

posing open-ended questions to students at strategic places in a text and encouraging students to build on, and challenge, ideas put forth by their peers. Teachers further cultivate students' comprehension skills by asking students to analyze and interrogate ideas, explain their thinking and reasoning, and back their claims with appropriate evidence (Goldman, Snow, & Vaughn, 2016; Langer, 1995; Murphy et al., 2009; Nystrand, 2006). Students, in turn, respond in longer ways, explain their reasoning with evidence, and connect their ideas to the thinking expressed by their classmates.

Dialogic text discussions are a stark contrast to traditional forms of classroom discourse that have characterized reading instruction in the large majority of classrooms (Applebee, Langer, Nystrand, & Gamoran 2003; Sedova, Salamounova, & Svaricek, 2014). In typical classroom discussions, teachers pose fact-based questions with a predetermined right answer and evaluate each student's response (Mehan, 1979). This pattern of discussion, often referred to as Initiate-Respond-Evaluate (IRE), is teacher-centered in that the teacher determines who speaks, what is talked about, and what stands as a correct response. Students respond to the teacher rather than one another and rarely explain their reasoning or provide extensive evidence for their ideas. Unsurprisingly, given the distance between classroom discourse that nurtures higher level reading comprehension skills and typical practice, national assessments have shown that only about a third of fourth-grade students are able to integrate information, infer meaning, and draw conclusions about what they read (National Center for Education Statistics, 2017).

Because dialogic text discussions are fundamentally different from traditional practice, transforming classroom discourse patterns is often a challenging endeavor for teachers (Kucan, 2009). Research on dialogic instruction more broadly has suggested that whereas teachers can readily incorporate new forms of questions in their discussions, responding to students in ways that expand their thinking and reasoning is a very difficult skill to master (Franke et al., 2009; Lefstein, Snell, & Israeli, 2015). Learning how to respond productively to students requires not only knowledge of a range of possible talk moves (questions and rejoinders) but also, more importantly, the ability to anticipate how a particular talk move is likely to shape the trajectory of students' thinking. This is important because simply incorporating new talk moves (e.g., asking open-ended questions) is unlikely to meaningfully increase students' discussion participation and reading comprehension skills (Boyd & Markarian, 2015; Lefstein et al., 2015).

Although substantively shifting patterns of classroom talk to be more dialogic is challenging, such transformation is possible. Research has suggested that professional development interventions that combine formal study (e.g., workshops, classes) to develop teachers' understanding of an instructional model, paired with coach-guided reflection to implement new forms of instruction, can be an effective strategy for increasing text discussion quality (Chinn, Anderson, & Waggoner, 2001; Matsumura, Garnier, & Spybrook, 2012, 2013; Murphy et al., 2018; Sedova, Sedlacek, & Svaricek, 2016). Although this research base provides important insight on the components of professional development that increase the quality of text discussions, what is less understood is how and why the activity settings that comprise these interventions contribute to change in teachers' practice. Indeed, teachers' professional learning in practice-based interventions more generally has been undertheorized (Munter & Correnti, 2017; Thompson, Windschitl, & Braaten, 2013). The varied approaches to professional development have created limited consensus about how teachers learn and, therefore, how professional development influences practice (Kennedy, 2016), and theories of how teachers appropriate conceptual and pedagogical tools suggest varied and highly individualized developmental paths based on a number of factors, including teachers' identity, access to tools, and context (Grossman, Smagorinsky, & Valencia, 1999).

In the present study, we took up the call for more theoretically driven investigations by proposing and testing a theory of change for describing how teachers' adaptive expertise (Hatano & Inagaki, 1986) for facilitating dialogic text discussions develops over time through participation in an online literacy-coaching program, Online Content-Focused Coaching. Specifically, we investigated the effect of the intervention on students' reading skills and text discussion quality and explored the contribution of the workshop and coaching components of the intervention to the development of teachers' facilitation skills, students' discussion participation, and students' reading achievement.

Developing Teachers' Adaptive Expertise at Facilitating Dialogic Text Discussions

Adaptive expertise is the ability to apply knowledge flexibly and strategically to meet particular goals (Hatano & Inagaki, 1986). Teachers with such expertise within specific instructional domains possess highly developed, well-articulated schemata or mental models[1]—that is, dynamic knowledge structures—that integrate related concepts based on an individual's prior knowledge, beliefs, experiences, and expectations (Anderson, 1984; Johnson-Laird, 1980; Spillane, Reiser, & Reimer, 2002). Teachers' mental models for instruction serve as a lens through which they make sense of their classroom interactions, including what they notice and the meaning they ascribe

to the events they notice, and are invoked to make predictions, guide decision making, and resolve ambiguities (Spillane et al., 2002; Taylor & Crocker, 1981). Well-developed mental models for teachers' pedagogy enable them to effectively process new information and draw on an integrated base of knowledge (i.e., that connects procedural, conceptual, and experiential knowledge) to make informed pedagogical decisions.

Adaptive expertise in the context of dialogic text discussions would, therefore, involve both the ability to create dialogic space for students and the facilitation skills to nurture students' emergent thinking and develop rigorous and coherent lines of inquiry. Hence, to become adept experts in dialogic text discussion, teachers must not only understand how to plan their discussions in ways that allow for student ideas and voices to proliferate but also have the skills to notice the substance of students' ideas and adapt their instruction (i.e., make pedagogical decisions) based on their interpretation of students' thinking in the moment.

Importantly, teachers' learning of new instructional concepts and practices is filtered through their existing mental models, the contours of which are typically defined as a transmission or teacher-centered model of instruction (e.g., traditional forms of classroom discourse). Hence, when teachers are introduced to new instructional models that run counter to their existing conceptions of what effective instruction looks like, they often add in new instructional concepts and practices but are unlikely to meaningfully alter their existing understandings and practices in substantive ways (Sedova et al., 2016; Spillane et al., 2002). For example, as described earlier, because most teachers model their understanding of effective classroom discussion by traditional, IRE standards of practice, they may incorporate certain features of dialogic instruction (e.g., ask open-ended questions) without substantively changing their discourse patterns (e.g., ask low-level literal follow-up questions, collect student ideas absent any attempt to explore their thinking; Lefstein et al., 2015).

This suggests that to substantively change teachers' discussion practices, and for those discussion patterns to be sustained over time, professional learning activities must also address teachers' existing mental models for effective classroom text discussion. Depending on the distance between teachers' existing mental models and dialogic discussion, significant restructuring may be required to avoid superficial adoption of new practices (Spillane et al., 2002; Taylor & Crocker, 1981).

Online Content-Focused Coaching

Online Content-Focused Coaching is based on a literacy-coaching model developed at the University of Pittsburgh's Institute for Learning (Matsumura et al., 2012, 2013; Matsumura, Garnier, Correnti, Bickel, & Junker, 2010). To make the program accessible to schools that might not have access to full-time coaches, an online application of the program was developed that consists of two components: a six-week online workshop followed by multiple cycles of remote coaching around teachers' video-recorded class discussions.

Online Workshop

The content of Online Content-Focused Coaching centers on Questioning the Author (Beck & McKeown, 2006) and Accountable Talk (Michaels, O'Connor, & Resnick, 2008) techniques. Questioning the Author draws on cognitive science research that has characterized text comprehension as an active and inferential process of building a mental representation of a text (Kintsch & van Dijk, 1978). Accountable Talk draws on sociocultural theory (Vygotsky, 1934/1986) and research in the learning sciences (Bransford, Brown, & Cocking, 2000; Greeno, Collins, & Resnick, 1996) and emphasizes the importance of participants building on the ideas of others, making logical connections between ideas, drawing reasonable conclusions, and making explicit the evidence behind claims. To plan for text discussions, teachers learn how to choose texts in their curriculum that contain sufficient grist (i.e., complexity in theme, plot, or language) to support rigorous discussions (Beck & McKeown, 2001). Teachers also learn how to develop questions that support students to extend meaning from texts and how to select appropriate stopping places in a text for posing questions. These can be places in texts where the authors do not express themselves clearly, use vocabulary that is likely to be unfamiliar to students, or introduce an idea or critical event that is ripe for students' exploration.

Workshop Activities

The design of the workshop activities draws on cognitive apprenticeship theory that emphasizes four key activities situated in authentic practice to build conceptual knowledge and skills: modeling, guided practice (scaffolding), fading, and coaching (Collins, 2006; Collins, Brown, & Holum, 1991). The online workshop focuses on the first three activities with the goal of introducing and developing teachers' mental models for enacting dialogic text discussions, with the coaching activity represented in the second component of the intervention.

Modeling

The workshop begins with teachers reading and watching webinars to build their understanding of the role of classroom talk in students' reading comprehension. Teachers then study videos of expert text discussions, and annotated lesson transcripts and lesson plans that make target practices (e.g., talk moves) explicit. Importantly, these annotations explain the reasoning behind an expert teacher's pedagogical choices, such as why a particular question was asked.

Teachers also compare their responses (e.g., their analysis of a lesson transcript) with an expert analysis and reflect on how their choices might have differed from the expert's choices. This activity is intended to develop teachers' capacity to notice key features of instruction (Sherin & van Es, 2009), and support conceptual understanding for why a talk move might accomplish a particular goal within a lesson.

Guided Practice

Teachers in the workshop also engage in guided practice to participate in the steps of planning for dialogic text discussions. Scaffolds are provided to teachers in the form of course tools. To help teachers choose texts from their curriculum that would support a rich discussion, for example, teachers are provided with a rubric for analyzing text complexity, and a lesson-planning template. Initially, teachers apply these tools to texts that are part of the workshop, and use one of these planned lessons as the basis for their first individual online coaching session. Over the coaching phase of the intervention, teachers plan lessons based on texts from their district's curriculum to support their application of Questioning the Author and Accountable Talk in their teaching context.

Online Coaching

Upon completion of the workshop, teachers participate in cycles of remote coaching ($N = 4$ or 5 cycles) with a coach from the University of Pittsburgh's Institute for Learning, beginning with a lesson they planned in the workshop. To guide coaching conversations, teachers are provided with a document that summarizes the content learned in the workshop—the Framework for Effective Text Discussions (see Appendix A)—that makes explicit the teacher and student talk moves that are characteristic of Questioning the Author and Accountable Talk. Each coaching cycle comprises three phases: a prelesson conference, a written reflection on a video-recorded lesson, and a postlesson conference.

Prelesson Conference

A coaching cycle begins with a phone conference to discuss the lesson plan for the classroom text discussion, with a focus on supporting teachers to see how features of their plan, such as stopping points and questions, are likely to shape students' thinking about a text. The coach and teacher then collaboratively determine the teacher's goals for student learning and pedagogy keyed to the dimensions of the Framework for Effective Text Discussions.

Written Reflection on a Video-Recorded Lesson

Teachers upload a video of their planned classroom discussion (approximately 30 minutes long). The coach then identifies three short segments (two- to three-minute clips) keyed to the established pedagogical goals to reflect on with teachers in the postlesson conference. The coach uploads the short video segments to an online coaching interface and poses reflective comments and questions for the teacher to respond to in writing within the online interface.

Postlesson Conference

A coaching cycle ends with an individual postlesson phone conference (approximately 30 minutes long), in which the coach and teacher watch the lesson clips together (each at their own computers) and compare the interactions in the lesson with the teacher and student talk moves represented in the Framework for Effective Text Discussions. The purpose for reflecting in this way is to focus the coaching conversation squarely on evidence of the instructional models (Questioning the Author and Accountable Talk) instantiated in a teacher's practice, as well as to surface discrepancies between the interactions shown in the teacher's video-recorded class discussion and targeted pedagogical principles and goals.

The coach supports the development of teachers' skills for facilitating dialogic text discussions when reflecting on the video-recorded lessons in several ways. First, the coach encourages teachers to interpret and explore explanations for discussion interactions and discourages evaluation or judgment of these events. Second, the coach guides teachers to notice and analyze evidence of student thinking to surface discrepancies between a teacher's pedagogical goal and what transpired in a lesson. For example, a coach might support a teacher to notice that students were responding in short, fact-based ways. In the process of understanding why this was the case, a coach might then support the teacher to see that he or she was tending to follow open-ended questions with closed-ended questions that constrained students' thinking, thus undermining his or her pedagogical goal of supporting rigorous thinking. Finally, the coach will, as needed, clarify concepts, challenge teachers' thinking, and invite teachers to consider how the use of another talk move in that same situation might have impacted students' sense-making opportunities. This latter practice of mentally trying out different talk moves is especially important because it helps develop teachers' conceptual understanding of why and how talk moves shape student thinking.

Theory of Teachers' Professional Learning and Change in Classroom Text Discussions

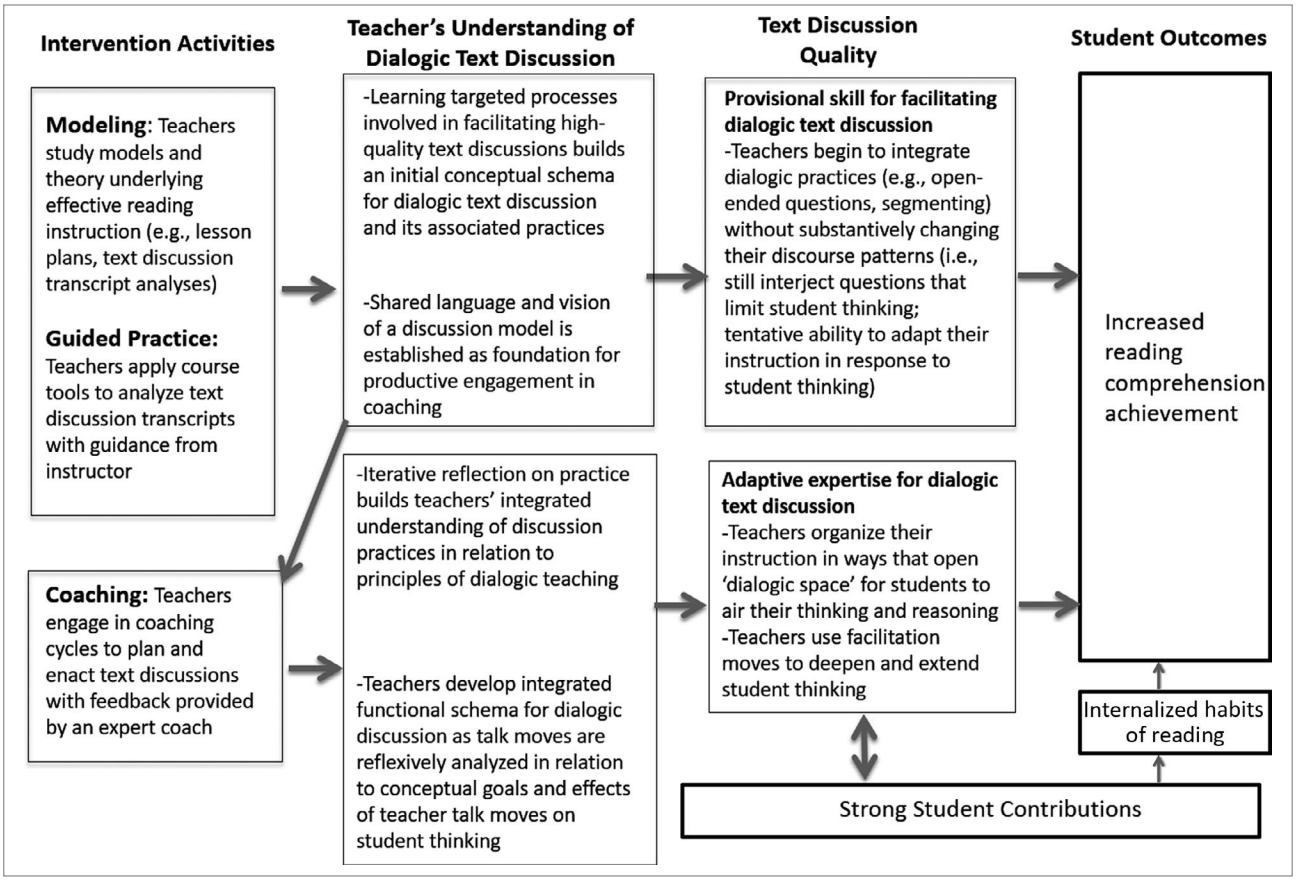
Our theory of a process of change with respect to how Online Content-Focused Coaching develops teachers' adaptive expertise is based on the assumption that

transformation in classroom text discussions from traditional IRE to dialogic requires change in teachers' conceptions of the form and purpose of their talk moves (initiating questions and rejoinders). Conceptual change can be inspired from something as simple as a belief change, but when knowledge is misconceived or flawed, conceptual change is more complex and difficult because it requires either a mental model shift or recategorization of knowledge (Chi, 2008). Thus, conceptual change often requires not only additional knowledge (e.g., the acquisition of new talk moves) but also resolution of conflicting beliefs, misunderstandings, or values to promote lasting change in behavior patterns. This is critical because one way humans process knowledge efficiently is through heuristics that promote assimilation for most types of received knowledge (Gregoire, 2003). Thus, like all learners, when teachers are introduced to new ideas, they are more likely to assimilate what is learned rather than seek to accommodate such knowledge, leading to surface-level or scant changes in practice.

Resolving the conflict between prior and new knowledge involves at least four prerequisites: Learners must become dissatisfied with the conflict between prior conceptions and new beliefs and practices, and learners must perceive new knowledge (e.g., a new conception of their discussion practice) as intelligible, as plausible, and as productive (Posner et al., 1982, as cited in Gafoor & Akhilesh, 2010). Creating the conditions for a mental model shift thus requires motivation to change on the part of learners (i.e., buy-in, the establishment of new goals) and the opportunity to become aware of the specific ways that current practices thwart the acquisition of these new goals. Given the complexity of such shifts in teachers' mental models, resolving prior conceptions about one's own teaching practice may require repeated examinations of the same conflict after dissatisfaction with a prior conception has already been identified. Thus, our developmental hypotheses suggest that evidence for conceptual change would not be seen immediately, but rather over time.

In the context of our intervention, Figure 1 describes how the components of Online Content-Focused Coaching intersect with this theoretical orientation.[2] Drawing on the work of P.A. Alexander (2003, 2004), we propose a three-stage sequence in dialogic text discussion quality from traditional IRE (before intervention) to provisional skill for eliciting student thinking and facilitating students' ideas (typically following the workshop) to

FIGURE 1
Theory of a Process of Change for Leading Text Discussions



adaptive expertise for dialogic text discussion (after a varying number of cumulative video-coaching cycles). Specifically, during the workshop, opportunities are provided to teachers to learn the subskills involved in planning for and enacting dialogic text discussions to support or establish their initial (or emergent) mental model for facilitating dialogic text discussions. Teachers were exposed to this new knowledge, specifically, what dialogic text discussions look like (e.g., through study of expert models based in authentic classroom practice) and, as importantly, why these discussions build students' reading comprehension skills (i.e., the purpose of new talk moves). This is important for helping teachers understand dialogic practices as both plausible and productive, thus establishing motivation to try out new practices in their classroom.

We expect, however, that the new knowledge acquired through the workshop would coexist with teachers' prior beliefs and practice. In other words, we did not expect at this phase of the intervention to see a significant shift in teachers' beliefs or practice. Instead, we expected that teachers would adopt open-ended questions and other talk moves that resemble elements of dialogic discussion (e.g., pressing students to explain their thinking), with little or no change in the quality of students' discussion participation. In other words, teachers in this phase of the intervention would not have yet developed the robust and well-integrated mental model needed to guide highly skilled pedagogical choices required to optimally respond to and facilitate students' own thinking and reasoning, resulting in a surface-level implementation of dialogic text discussions.

In the coaching phase of the intervention, through engaging teachers reflexively[3] around their classroom discussions, we expected that teachers would integrate the subskills and knowledge needed to use talk moves strategically in pursuit of desired goals (i.e., cultivate a more fully developed mental model). This shift would occur through the coach assisting teachers to see the conflict between the interactions that transpired in their video-recorded lessons and goals for dialogic text discussion practice (as instantiated in the Framework for Effective Text Discussions). More specifically, because mental models can be viewed as bundles of cause–effect associations (Wilke & Losh, 2012), we expected that the atomistic work of revising mental models would operate at the level of a teacher examining how particular teaching moves in discussion resulted in (un)desirable student responses (e.g., how a teacher's following an open-ended question with a closed-ended question resulted in factual recall on the part of students instead of higher level thinking about a text). The coach assists this process of conceptual change by drawing attention to specific interactions in the vide-recordoed lesson (e.g., a question and response), as well as by filling in the gaps in teachers' knowledge as needed (e.g., clarifying the difference between open-ended and partially open-ended questions, considering how alternative talk moves might have differentially impacted the trajectory of student thinking). Thus, through coaching, we expect to see a more complete shift in teachers' mental model for dialogic text discussions and an attendant increase in teachers' ability to choose more optimal talk moves in the moment of a discussion to elicit and extend student thinking (see, e.g., Franke et al., 2009; Thompson et al., 2013). This development in teachers' adaptive expertise, the ability to use talk moves more productively, would be shown in an increase in the amount and quality of students' discussion participation (McNeill & Pimentel, 2010; Soter et al., 2008). Growth in the rigor and interactivity in classroom discussions, in turn, would increase students' reading skills, partly because students will internalize habits of reasoning acquired through their participation in such text discussions (see Murphy et al., 2009; Nystrand, 2006; Wilkinson et al., 2015).

The Present Study

We seek initial evidence for the efficacy of Online Content-Focused Coaching, as well as evidence for our theory of a process of change in a small-scale randomized trial ($n = 31$ teachers: eight treated and 23 controls). We begin by examining effects of Online Content-Focused Coaching on students' reading achievement. We then examine the effect of Online Content-Focused Coaching on classroom text discussion practices to investigate differences between treated and control teachers in their growth curves, while adjusting for initial starting points in classroom text discussion quality. We then examine trends in discussion patterns and survey responses in relation to our hypotheses around the development of teachers' adaptive expertise. We seek evidence in support of our theory of teachers' development, that is, demonstrating a three-stage process of teaching development toward adaptive expertise. Finally, we provide an example of change in text-discussions in one teacher's classroom. Our first three formalized research questions follow:

1. What is the effect of Online Content-Focused Coaching on students' reading achievement?
2. What is the effect of Online Content-Focused Coaching on classroom text discussion quality (teachers' facilitation and students' discussion participation)?
3. How do text discussion quality and teachers' self-reported beliefs and practices develop across different phases of the intervention (i.e., in the workshop and coaching)?

Evidence from all of these research questions contributes to a validity argument for the efficacy of Online Content-Focused Coaching, from predictive validity for the intervention on changes in student achievement, to main effects of Online Content-Focused Coaching on teaching outcomes and teachers' beliefs, to tests of evidence for our theorized process of change. Meanwhile, both changes in beliefs and evidence for stages of improvement in text discussion quality provide inductive support for our theorized process of teaching change because conceptual change is thought to involve challenges to prior beliefs or dispositions (Chi, 2008; Gregoire, 2003; Wilke & Losh, 2012).

Finally, we explored whether we observed an association between improvements in text discussion quality and classroom-level changes in students' reading achievement. In this analysis, we investigated the theorized relation in the last two columns of Figure 1, between text discussion quality (including students' strong contributions) and students' reading achievement. This led to our fourth research question:

4. What is the relation between growth in classroom text discussion quality and growth in students' reading achievement?

Method

Study Design and Randomization

We employed a pretest–posttest longitudinal cluster-randomized design (see, e.g., Shadish, Cook, & Campbell, 2002) to test effects of Online Content-Focused Coaching on student and teacher outcomes. Schools were chosen as the unit of randomization to minimize the potential for diffusion of treatment threats to our design (McMillan, 2007). In the spring preceding the September start date of our study, we presented our research plan to all of the elementary school principals in a medium-sized district in Texas. Fourth- and fifth-grade teachers in the schools whose principals agreed to support the study activities were contacted by email and invited to participate in the study. Teachers ($N = 62$) who signed on for the study indicated their willingness to participate regardless of their assignment to the treatment or control condition. We used stratified random sampling techniques to account for potential imbalance that might result from simple random sampling (Ivers et al., 2012; Lohr, Schochet, & Sanders, 2014). Notably, because schools varied markedly in the percentage of English learners, we grouped schools into roughly three strata: (1) low—fewer than 20% English learners ($n = 7$ schools; 21 teachers), (2) medium—between 20% and 50% English learners ($n = 12$ schools; 28 teachers), and (3) high—greater than 50% English learners ($n = 8$ schools; 13 teachers). A check of random assignment showed that it produced roughly equivalent groups of schools[4] on a number of student demographics, including prior reading and math achievement, school size, percentage of educationally disadvantaged students, percentage of at-risk students, and percentage of mobile students. A coin toss was used to assign the groups of schools to conditions.

Attrition

To meet the study's single-year timeline, treatment teachers began the workshop in late September, with coaching beginning in December or January. Teachers in both conditions completed a survey and video recorded their class discussion early in the school year, and treatment teachers then immediately participated in the workshop (a commitment of three to four hours of time per week). We note this as important because our study design, particularly the confluence of an intensive workshop in the midst of data collection responsibilities, may have factored into attrition rates of treated teachers at the beginning of the year, as they were balancing work commitments and getting to know their classroom of students. Although teachers in both conditions received the same compensation for completing the research activities ($1,500), control teachers were not offered the workshop until the end of the school year, after accountability testing.

As detailed in Appendix B, differential attrition occurred, with a greater number of treatment teachers attriting from the study than control teachers. First, only 21 teachers took the initial pre-survey indicating their intention to begin the treatment as opposed to 29 teachers in the control group, suggesting greater attrition over the summer. Six of the 21 treatment teachers attrited before the workshop began, mostly for personal reasons. Another seven teachers did not complete the workshop (i.e., 47% attrited near the beginning of the workshop). Our understanding is that the differential attrition was due mostly to the confluence of study demands (workshop plus data collection activities) placed on teachers at the same time in the fall. In contrast, attrition from the control group for data collection alone (21%) and during the workshop in the spring (9%) were both far less dramatic when they occurred independently. In the future, we would choose to conduct the workshop in the spring (or summer) prior to a full year of coaching. For analytic purposes, however, we found limited evidence for statistically significant between-group differences in *t*-tests for baseline equivalence (missing data analyses are detailed in Table B2 in Appendix B). In addition, we took steps to mitigate any observed preexisting group differences during the analytic phase to better support our inferences.

Demographics of Our Analytic Sample

Of the 31 teachers who formed the analytic sample ($n = 8$ treated and 23 controls), 18 taught fourth grade, and 13

taught fifth grade. Teachers ranged in years of teaching experience, with a fairly normal distribution around the mean of 13 years of experience. Teachers also varied in their exposure to English language arts methods courses in their teacher preparation programs, with eight teachers (roughly 25%) having taken fewer than four courses and the others reporting from four to more than 10 courses. Most teachers (82%) held regular teaching certification, and all held a bachelor's degree. Forty-three percent also held a master's degree. The majority of students (61%) in our study schools were from low-income families. The largest proportion of students were identified as Latinx (73%), followed by Caucasian (15%), African American (7%), multiracial (4%), and Asian or Pacific Islander (1%).

Measures and Procedures

State of Texas Assessment of Academic Readiness (STAAR)

The STAAR standardized test is the state assessment administered to students in the elementary and secondary grades. In reading, the test involves multiple-choice and short-answer responses to text passages. The reading test comprises three subscales: understanding/analysis across genres, understanding/analysis of literary texts, and understanding/analysis of informational texts. The overall reading test (combined subscales) has a high level of reliability in third (Cronbach's $\alpha = .89$), fourth ($\alpha = .91$), and fifth grades ($\alpha = .91$; Texas Education Agency, 2015).

Video-Recorded Classroom Text Discussions

Teachers in the treatment condition video recorded themselves leading text discussions in their classrooms on the following schedule: at baseline (before beginning the workshop in September), at the end of the workshop (December or January), and for each of their subsequent coaching sessions from February through May ($n = 44$ text discussion videos, $\mu = 5.86$). Teachers in the control condition video recorded themselves leading text discussions in early September, in December or January, and at the end of April prior to beginning the workshop ($n = 69$ text discussion videos, $\mu = 3.0$). Two approaches were used to assess the quality of text discussions, described in turn next.

Instructional Quality Assessment (IQA)

The first approach rated discussion quality using an observation measure that past research showed statistically significantly predicts gains in students' reading scores (Matsumura et al., 2013; Matsumura, Garnier, Slater, & Boston, 2008). The IQA comprises eight dimensions, plus we developed two additional dimensions, segmenting the text and constructing the gist, to better represent some of the primary features of Questioning the Author. Each dimension was rated on a 4-point scale (1 = *poor*; 4 = *excellent*) with the exception of the rigor of the text dimension, which was rated on a 3-point scale (1 = *low*; 3 = *high*). For example, ratings for the teacher links student contributions dimension range from 1 (the teacher does not make any effort to link or revoice students' contributions) to 4 (the teacher consistently, that is, three or more times during the lesson, connects speakers' contributions to one another and shows how ideas/positions shared during the discussion relate to teach others; e.g., "What I hear you saying is that the character has changed from the beginning to the end of the book, and that position is similar to Ana's [all names are pseudonyms] idea that the character has matured throughout the book").

Reliability of IQA Scores

To score the videos on the IQA, two raters, each with a PhD and experience in education research, were hired and trained by a member of the research team. Training took place over a three-day period and focused on studying the codebook for text discussions, coding short segments of video, and learning to use Studiocode software (Vigital, 2015) to code the video-recorded lessons. All videos were deidentified, and the order in which they were scored was randomized so raters were blind to the teacher, condition, and when the video was obtained during the academic year. To assess inter-rater reliability of measures from our video-recorded text discussions, 32 videos (26%) were double-coded at the dimension level by the raters. Intraclass correlation coefficient (ICC) estimates were calculated via SPSS version 26 using a single-measure ($k = 2$), absolute agreement, two-way random effects model. The top half of Table 1 provides ICCs at the dimension level (ICC ranges = .89–.98). Similar to prior work (Matsumura, Correnti, Walsh, Bickel, & Zook-Howell, 2019), we engaged in data reduction during analysis to create three composite scores of the individual IQA dimensions based on theoretical groupings aligned with Online-Content-Focused Coaching; (1) choice of a rigorous text and implementation of Questioning the Author, (2) teachers' use of Accountable Talk moves to facilitate text discussions, and (3) rigorous student contributions. ICCs for these three composite constructs (ICC range = .94–.96) revealed excellent agreement between raters (Koo & Li, 2016).

Exhaustive Coding of Text Discussions

Our second approach was to conduct exhaustive coding of the same video-recorded lessons to investigate changes in teachers' facilitation moves and students' contributions from a developmental perspective (see Appendix C for an excerpt of the codebook, sampling only the codes we used analytically for demonstrating the process of change).

TABLE 1
Inter-Rater Reliability for Coding Video -Recorded Text Discussions

Dimensions of Instructional Quality Assessment text discussion quality	Intraclass correlation coefficient
Rigorous text and Questioning the Author implementation composite[a]	
Rigor of text[a]	.89
Segmenting the text	.98
Guidance toward constructing the gist	.95
Teacher facilitation moves composite	
Teacher links student contributions	.92
Teacher presses students to support assertions	.90
Developing community	.93
Teacher poses cognitively demanding questions	.92
Rigor of student contributions composite	
Students link their contributions to other students'	.96
Students provide text-based evidence for assertions	.89
Students provide extended explanations for assertions	.89
Measures for examining process changes	**Intraclass correlation coefficient**
Partially open-ended questions—transitional	.57
Open-ended gist questions—aspirational	.85
Uptake-literal questions—transitional	.73
Uptake questions—aspirational	.67
Weak student contributions—transitional (3 items)	.78
Strong student contributions—aspirational (3 items)	.82

[a]We adjusted the rigor of text score from its original metric (1–3) to the same metric as the rest of the items (1–4) before calculating the mean and checking inter-rater reliability.

Specifically, we coded every teacher turn (22 unique codes total, including initiating and rejoinder moves) and student turn (six unique codes total) in the video-recorded text discussion, building off of a prior coding scheme (Correnti et al., 2015). Analytically, to understand change in teachers' text-based discussions over the year, we sampled codes that captured talk moves that we defined as aspirational, that is, targeted teacher discussion moves (launches and rejoinders) that we see as elemental for dialogic discussions (e.g., asking an open-ended question to construct the gist of a text, uptake of a student's idea to deepen and/or extend discussion). Importantly, we also measured codes that represent transitional moves, that is, talk moves that teachers might employ as they begin to change deep-seated patterns of discourse in attempts to move toward dialogic talk moves (e.g., asking a partially open-ended question or uptake of a student's idea in a literal fashion, thereby missing the opportunity to extend an idea). The categorization of transitional and aspirational codes allowed us to run preliminary tests for whether we saw evidence of a three-stage process of development. We anticipated immediate influence of the workshop on transitional moves, followed by a switch from transitional moves to aspirational moves during coaching.

Reliability of Exhaustive Coding Measures

Given that our analyses will examine growth of specific talk moves, we also examined inter-rater reliability of certain individual codes from our exhaustive coding. To assess inter-rater reliability of individual codes from the video-recorded text discussions, we used the same 32 videos (26%) that were double-coded at the turn level by expert raters. We aggregated across all turns in a lesson for each individual code. We then calculated the ICC for the frequency counts for each individual code examined in subsequent growth models. The bottom half of Table 1 provides ICCs for four individual codes for teacher talk moves that were explored to understand change in frequency of these moves over the year. These ICCs indicate moderate to good inter-rater reliability for counts of these

individual codes across the 32 videos. Additionally, we calculated the ICCs by comparing the counts of weak and strong student contributions from both raters on the 32 videos, and both ICCs indicate good reliability (Koo & Li, 2016).

Teacher Surveys

Teachers in both conditions were surveyed at the beginning (baseline) and end of the academic year. The surveys focus on teachers' beliefs and self-reported practices for planning and facilitating dialogic text discussions. Items included desirable practices overlapping with Questioning the Author and Accountable Talk (e.g., endorsement of open-ended questions), as well as items that represented IRE patterns of discourse (e.g., endorsement of literal questions for building comprehension skills). We engaged in a set of data reduction activities to group items, resulting in six constructs. Table D1 in Appendix D provides the items within each factor, as well as the Cronbach's alpha for each construct (α range = .60–.87).

Analyses

Effects on Students' Reading Achievement

Using student scores on the state standardized test, we examined models to explore whether there was a main effect of the treatment on students' reading achievement. We first examined hierarchical covariate-adjusted models while adjusting for student and classroom covariates. Using HLM version 7.03, we examined models for our analytic sample to determine estimates of the treatment effect in naive statistical models. These models nested students within teachers, with the treatment at the teacher level of the model. One advantage of these models was that they used as much data as possible, because HLM allows for missing data at level 1. We followed up these analyses with different strategies for estimating the treatment effect when a study has experienced differential attrition. We used a tobit model to understand the sensitivity of our findings to truncated distributions or serious departures from normality. We also used selection-corrected regression models (Heckman's method; Heckman, 1979) to understand whether treatment estimates were sensitive to factors that were related to attrition. These models account for potential attrition bias that occurred as the result of discontinuing participation in the treatment. We present results from all models to understand the sensitivity of model-based estimates to the choice of method.

Effects on Text Discussion Quality

In our first examination of the data, we compared the two primary endpoints for treated and control teachers. Thus, we examined paired-samples *t*-tests for each IQA dimension and for each of the three composite scores. These analyses provided one way to compare effects for treated and control teachers by measuring growth from baseline to the last measured classroom text discussion.

To understand growth trajectories for each group of teachers and to better describe overall patterns of growth, we examined the data using hierarchical linear models to identify the best functional form for change on each of the IQA composite scores over time, that is, choosing rigorous texts and implementing elements of Questioning the Author, teacher press and facilitation moves, and rigor of student contributions (see Appendix E for model details). These models use our repeated-measures video data to show how patterns of text discussion quality grew over time, estimate the variability in growth between teachers, and understand the reliability of the growth estimates (as a proportion of between-teacher variability relative to within-teacher variability for the IQA composite scores). Furthermore, when an indicator for treatment is added to the prediction model, the models allow for a direct comparison between model-based rates of growth for the treated teachers as compared with the control teachers.

To further investigate group differences in growth and determine an effect size estimate, we examined a series of different statistical models to explore whether the treatment effect was sensitive to model selection. By running these models, we were trying to alleviate our concerns that findings could be due to the differential attrition we observed and/or to the treated group of teachers beginning the study higher than the control teachers on their baseline IQA composite scores for the construct measuring teacher press and facilitation moves (μ_{tx} = 2.19; μ_{c} = 1.67; p = .008; Hedges' g = 1.13).[5] Consequently, we examined naive regression models with statistical adjustments, a model with a linear adjustment for a propensity-for-attrition score, a model with propensity strata included as predictors and finally, a differences-in-differences model.[6] Ultimately, we relied on the differences-in-differences model for our calculation of effect sizes. All models led to a similar inference about a treatment effect, but the differences-in-differences model yielded the most conservative estimates because it relies primarily on the condition-by-time interaction, while accounting for any pretreatment differences between groups, as well as a main effect of time.

Patterns of Growth in Teacher Outcomes Across Phases of Online Content-Focused Coaching

We also examined piecewise hierarchical linear growth models (Raudenbush & Bryk, 2002) to describe different patterns of growth in aspirational and transitional codes for treated and control teachers. These models provide

linear estimates of growth for each interval, that is, from baseline to end of workshop (i.e., workshop interval) and from end of workshop to each teacher's last video (i.e., coaching interval). We ran models for each of our six codes—three aspirational and three transitional—to identify patterns of change (see Appendix F for model descriptions).

To follow up on these six univariate analyses, we engaged in data reduction to generate an outcome across all items. First, we summed across transitional items (i.e., proportion of all teacher turns that were partially open-ended launches, uptake-literal rejoinders, weak student contributions[7]), and we also summed across aspirational items (i.e., proportion of all teacher turns that were open-ended gist launches, uptake questions that deepened and extended the discussion, strong student contributions). To investigate our theory of change, we created a ratio of the proportion of aspirational items divided by the proportion of transitional items. Our theory of a three-stage process of change predicted growth in this ratio during coaching as teachers developed more nuanced mental models. Stated differently, we expected that both the proportion of aspirational turns would increase and that, simultaneously, the proportion of transitional turns would decrease because as teachers moved toward adaptive expertise, they would have greater capacity to employ aspirational moves instead of transitional moves. Thus, we examined a piecewise linear growth model using the ratio as an outcome to generate evidence for the efficacy of coaching consistent with our theory of change. Finally, we provide short excerpts from text discussions at baseline, immediately after the workshop, and after four cycles of coaching as an illustrative example of the general trends from these quantitative analyses.

Evidence for Change in Self-Reported Beliefs and Practices

Similar to prior analyses for growth in discussion quality, we employed a differences-in-differences model to understand whether there were statistically significant differences in growth between treated and control group teachers in survey responses of their discussion beliefs and practices.[8] We hypothesized declines for treated teachers on three constructs from the surveys—teacher-led basic comprehension following IRE patterns, basic-skills comprehension instruction, and focus on non-text-based factors (independent reading and personal connections)—because our instructional model promotes dialogic practices that extend beyond these text-based comprehension activities. For the second set of three constructs—complex texts form the basis for rich text discussion, implementation of Questioning the Author, and students provide meaningful contributions to advance the text discussion—we hypothesized gains for our treated teachers because those beliefs and practices were aligned with the goals of our instructional models.

Association Between Changes in Text Discussion Quality and Changes in Students' Reading Achievement

To further explore the main effects of the treatment, we developed classroom-based estimates of teaching growth. We also developed a value-added measure of students' gains in reading achievement. Our assumption was that if students made statistically significant improvements in their reading achievement over the course of a single year, it might be correlated with the changes in teaching that we observed. To calculate value-added scores, we generated covariate-adjusted ordinary least squares residuals at the classroom level after accounting for prior achievement, student background characteristics, school strata, school demographics, and grade level. We used HLM version 7.03 (Raudenbush & Bryk, 2002) to explore the following covariate-adjusted mixed model:

$$Y_{ij} = \gamma_{00} + \gamma_{q0} X_{ij} + r_{ij} + \mu_{0j}$$

where r_{ij} and μ_{0j} are assumed to be independent normal residuals and where Y_{ij} is the reading achievement scale score (academic year 2016–2017) for student i nested in classroom j; $\boldsymbol{X}_{ij}$ is a vector consisting of the ij student's reading achievement scale score in the previous year (academic year 2015–2016) and student background characteristics, including gender, race, and designation as English learner, gifted, economically disadvantaged, or different levels of special education designation. Of primary interest in these models is μ_{0j} representing the teacher random effects or the teacher value-added after adjusting for the student covariates in the model, including prior achievement.

Next, to calculate a measure of teaching growth, we calculated a mean of the following measures: (a) estimates of the growth slopes from each of the three prior growth models of the video-based text discussions for each IQA composite score (without treatment as a covariate) and (b) gains in teachers' self-reported beliefs and practices from the survey from baseline to spring (i.e., the three constructs with hypothesized gains: complex texts form the basis for rich text discussion, implementation of Questioning the Author, and students provide meaningful contributions to advance the text discussion). We then examined the scatterplot, along with bivariate and partial correlations between our estimates of value-added achievement gains and teaching growth among treated and control teachers.

Findings

What Is the Effect of Online Content-Focused Coaching on Students' Reading Achievement?

Our estimates for a treatment effect in naive hierarchical linear models introduced our statistical controls at the student level, as well as the school strata and the grade level of the class. We also examined a number of school and classroom covariates, but given that none were statistically significant, we omitted them from the final prediction models. These models include 27 of the 31 teachers in our analytic sample.[9] Two teachers were omitted because we did not receive achievement records for students connected with those teachers, and two of the control teachers were omitted because they were from schools in the strata where no treated teachers completed the study. Tables 2 and 3 display the findings after adjusting for statistical controls for the differences between treated and control classrooms in predicted reading achievement scale scores. Table 2 shows a

TABLE 2
Bivariate Normal Hierarchical Regression Results in a Naive Model With Statistical Controls

Covariate	Coefficient	Standard error	Coefficient	Standard error
Intercept, γ_{00}	1,566.86	6.59	1,556.48	7.12
Treated, γ_{01}			35.02*	13.19
Stratum 1, γ_{02}	28.52*	14.69	27.25*	13.33
Grade 4, γ_{03}	18.52	18.24	23.42	16.55
Grade 5, γ_{04}	2.91	19.08	9.78	17.20
Prior reading achievement, γ_{10}	0.51***	0.06	0.51***	0.06
Prior math achievement, γ_{20}	0.15**	0.05	0.16**	0.05
Days absent, γ_{30}	−0.14	0.77	−0.15	0.77
Female, γ_{40}	19.73*	9.40	18.40*	9.38
Hispanic, γ_{50}	4.95	11.93	3.94	11.85
African American, γ_{60}	−10.84	17.55	−14.44	17.46
Native American, γ_{70}	14.80	29.12	14.57	28.86
Multiracial, γ_{80}	−14.67	21.06	−15.66	20.96
Asian, γ_{90}	−18.65	32.35	−19.30	32.21
At risk, γ_{100}	−25.05*	13.20	−27.56*	13.11
Special education, γ_{110}	−6.90	18.44	−6.11	18.33
Handicapped, γ_{120}	5.85	16.04	9.29	16.01
Economically disadvantaged, γ_{130}	−21.67*	10.35	−23.02*	10.29
Gifted, γ_{140}	61.19***	14.10	60.91***	13.80
English learner, γ_{150}	30.81	59.57	25.20	59.39
Limited English proficient, γ_{160}	−25.52	20.92	−21.12	20.73

Variance component	Variance	x^2	*p*	Variance	x^2	*p*
Between students within schools, *r*	9,015.90	43.89	.006	8,993.55	33.39	.056
Between schools, μ	539.92			322.06		
Variance between schools explained by treatment				40%		

*$p < .05$. **$p < .01$. ***$p < .001$.

TABLE 3
Sensitivity of Treatment Effect Findings on Achievement to Model Selection

Model	Reading scale score	
	Coefficient	Standard error
HLM naive model with controls (*n* = 467 students; 27 classrooms)	31.49*	13.55
Tobit model (*n* = 467 students; 27 classrooms)	29.16*	14.33
Heckman's selection method model (*n* = 403 students; 25 classrooms)	34.56*	15.40
HLM with a propensity score (*n* = 403 students; 25 classrooms)	38.66*	15.32
HLM with propensity score strata (*n* = 403 students; 25 classrooms)	38.90**	14.25

Note. HLM = hierarchical linear model. All models included covariates adjusting for student prior achievement, background characteristics, nesting of students in teachers, school strata, and grade level.
*$p < .05$. **$p < .01$.

statistically significant coefficient for the treatment and also how the addition of treatment to the model, by itself, explains 40% of the between-classroom variance in achievement scores.

The beta coefficients for the treatment effect across models shown in Table 3 reveal a consistent pattern, demonstrating that inferences about a treatment effect on student achievement are not sensitive to model selection. In particular, we discuss the Heckman model in Table B3 in Appendix B because it employs a regression with selection correction. The consistency of the findings from the Heckman model and all other statistical models suggests that differential attrition did not have a substantive influence on our estimates from the outcome analyses. As shown in Table 3 (and in greater detail in Table B3 in Appendix B), findings from the second stage of the Heckman model predict that students in classrooms of the treated teachers will score higher by 34.56 points (effect size [ES] = 0.24; $z = 2.24$; $p = .025$).

What Is the Effect of Online Content-Focused Coaching on Classroom Text Discussion Quality?

Instructional Quality Assessment Growth Estimates

In general, classroom text discussion quality improved for the treated group but remained the same for the control group. Results from the paired-samples *t*-tests, examining within-group growth, provide evidence of growth for the treated group on five of the 10 IQA dimensions by the end of the year but, more importantly, statistically significant improvement on each of the three IQA composite scores. Meanwhile, control teachers improved on just one of the 10 IQA dimensions (i.e., segmenting the text), and there was no improvement on any of the three IQA composite scores (see Table 4).

Although the paired-sample *t*-tests provide a snapshot of the differences in growth between treated and control teachers, they do not take advantage of the repeated-measures data (see, e.g., Rogosa, 1995; Singer & Willett, 2003) to describe teachers' trajectories for improvement in text discussion quality. Table 5 contains several noteworthy findings for the study of growth in text discussion quality on each of the three IQA composite scores. First, variance component estimates reveal a large change from the first conditional model (with classroom-level predictors) to the final conditional model (adding Online-Content-Focused Coaching as a predictor along with the same set of classroom-level predictors). The addition of treatment, by itself, explained a substantial proportion of the variance in growth (approximately 24% of the remaining variance in choice of rigorous text and implementation of Questioning the Author, approximately 90% of the remaining variance in teacher facilitation moves, and approximately 33% of the remaining variance in rigor of student contributions). Moreover, Online Content-Focused Coaching was the only statistically significant predictor of growth. Finally, it is notable that the measure of intra-individual change—sigma squared, usually conceived of as measurement error between timepoints—is quite small relative to the variance in growth. This suggests a relatively high degree of within-teacher consistency in IQA scores over time and resulted in reliable estimates of teaching growth in the fully unconditional models.

Figure 2 plots the growth estimates for each of the three IQA composite scores for the treated and control teachers. In each case, the best functional form for the growth trajectory was quadratic, indicating a rapid improvement on IQA measures (for treated teachers) followed by a tapering of growth (see Figure 2). One important feature of Figure 2 is the difference in growth trajectories on each of the IQA composite scores for the treated teachers versus the control teachers. Additionally, for the choice of rigorous text and implementation of Questioning the Author composite and

TABLE 4
Paired-Samples *t*-Tests for Change in Classroom Text Discussion Quality

	Treated (*n* = 8)		Control (*n* = 23)	
Dimension of the Instructional Quality Assessment	**Baseline preworkshop (September)**	**Last video-recorded coaching cycle (April or May)**	**Baseline preworkshop (September)**	**Last text discussion video (April)**
Rigor of text[‡]	2.25	2.75	2.09	2.09
Segmenting the text	3.38	3.63	3.39	3.91*
Guidance toward constructing the gist	1.75	3.50*	1.57	1.61
Rigorous text and Questioning the Author implementation composite[a]	**2.67**	**3.58***	**2.53**	**2.72**
Teacher links student contributions	2.13	2.63	1.30	1.35
Teacher presses students to support assertions	3.88	3.75	3.13	3.39
Developing community	1.63	2.88*	1.17	1.04
Teacher poses cognitively demanding questions	1.13	1.75	1.09	1.00
Teacher facilitation moves composite	**2.23**	**2.83***	**1.69**	**1.68**
Students link their contributions to other students'	1.63	2.38*	1.22	1.00
Students provide text-based evidence for assertions	2.38	4.00*	1.74	1.61
Students provide extended explanations for assertions	2.13	3.46*	1.87	1.74
Rigor of student contributions composite	**2.38**	**3.13***	**1.61**	**1.45**

[a]We adjusted the rigor of text score from its original metric (1–3) to the same metric as the rest of the items (1–4) before calculating the mean.
*$p < .05$.

for the rigorous student contributions composite, the estimated growth trajectory approaches the highest scale score of 4 and then plateaus.[10] Notably, control teachers demonstrated little to no improvement on these same measures over the same time interval. Additionally, we examined alternative models and found the treatment effect to be insensitive to model specifications.[11]

Differences-in-Differences Estimates

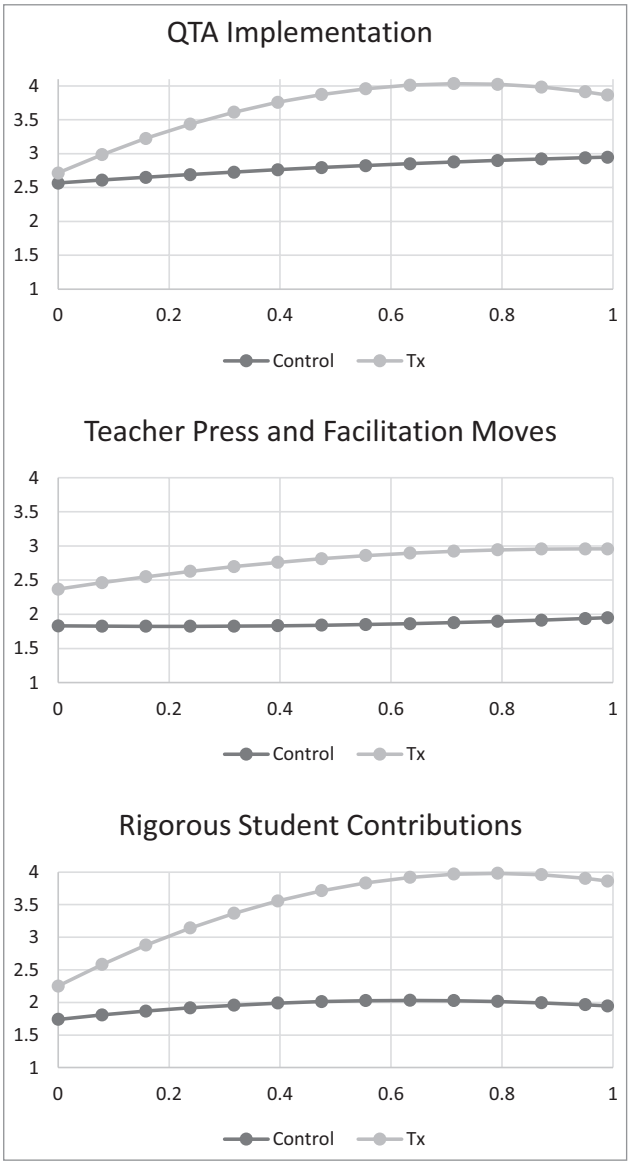
To determine effect size estimates, we examined a number of statistical models. Table 6 presents the estimates for the treatment effect on each of the IQA composite scores, revealing that the inference for differences between treated and control teachers is not sensitive to model selection. To calculate effect size estimates, we used estimates from the differences-in-differences model where the two endpoints for each of the teachers in the analytic sample were examined as a time-by-treatment interaction while including time and treatment as main effects in the analysis. These estimates are reported here to account for the noted difference in starting points between the treated and control groups on the IQA composite scores for teachers' implementation of Accountable Talk, and the trend of higher IQA scores on the other two IQA composites. Despite these being the most conservative estimates of the magnitude of the effect, we found effect sizes for the treatment on choice of rigorous texts and implementation of Questioning the Author (ES = 1.13, $p = .063$), on teachers' use of facilitation moves (ES = 1.14, $p = .007$), and on the rigor of students' contributions in discussions (ES = 1.21, $p = .002$). These findings are consistent with and extend prior work on face-to-face Content-Focused Coaching on the IQA[12] (Matsumura et al., 2012, 2013), as well as earlier pilot studies of Online Content-Focused Coaching, where patterns of improvement were quite similar (Matsumura et al., 2019).

TABLE 5
Hierarchical Linear Model Estimates for Growth on the Three Instructional Quality Assessment Composites

	Rigorous text and Questioning the Author implementation composite			Teacher facilitation moves composite			Rigor of student contributions composite		
Covariate	FUM coefficient (*SE*)	Model 1[a] coefficient (*SE*)	Model 2[a] coefficient (*SE*)	FUM coefficient (*SE*)	Model 1 coefficient (*SE*)	Model 2 coefficient (*SE*)	FUM coefficient (*SE*)	Model 1 coefficient (*SE*)	Model 2 coefficient (*SE*)
Intercept	2.56	2.57	2.57	1.81	1.82	1.83	1.73	1.82	1.74
Treatment			0.19			0.56**			0.66
Linear time estimate	0.71	0.65	0.53	0.10	0.08	−0.09	1.09	0.08	0.75
Treatment			2.75*			1.45†			2.01†
Quadratic time estimate	−0.33	−0.24	−0.13	−0.00	0.03	0.20	−0.91	0.03	−0.45
Treatment			−1.81			−1.08			−0.49
Variance components									
Intercept	0.36	0.32	0.32	0.10	0.10	0.05	0.80	0.72	0.66
Linear growth (percentage of reduction from the previous model)	6.00	6.33 (0%)	4.68 (26%)	1.03	0.50 (53%)	0.10 (80%)	7.15	4.92 (31%)	2.53 (49%)
Quadratic growth (percentage of reduction from the previous model)	3.64	3.63 (0%)	2.99 (18%)	0.67	0.37 (54%)	0.03 (92%)	4.64	4.21 (9%)	2.93 (30%)
Between timepoints	0.12	0.13	0.13	0.11	0.12	0.12	0.17	0.13	0.14

Note. FUM = fully unconditional model; *SE* = standard error.
[a]Model 1 accounts for classroom-level covariates such as prior year reading achievement, percentage of economically disadvantaged, average days absent, percentage of minority, grade level, percentage of gifted and talented, and percentage of limited English proficient. Model 2 contains the same covariates but adds assignment condition where treatment = 1 if treated, and treatment = 0 otherwise.
*$p < .05$. **$p < .01$. †$p < .1$.

FIGURE 2
Change Over Time in Instructional Quality Assessment Composites From Videos of Text Discussion Quality in Treatment and Control Classrooms



Note. QTA = Questioning the Author; Tx = treatment.

How Do Text Discussion Quality and Teachers' Self-Reported Beliefs and Practices Develop Across Different Phases of the Intervention?

Figure 3 provides visual representations of six piecewise linear growth estimates for change during the workshop interval (time from 0 to 0.33 on the *x*-axis) and during the coaching interval of Online Content-Focused Coaching (time from 0.33 to 1 on the *x*-axis). Figures 3a–c show the difference between treated and control teachers on transitional talk moves that we theorized would decline during coaching if teachers were learning to enact elements of dialogic discussion. Indeed, each of these figures shows a negative slope for Online Content-Focused Coaching teachers, especially in relation to the control teachers whose slopes remain relatively flat over the same time interval. Figures 3d–f show the difference between treated and control teachers on aspirational moves that we theorized would increase if teachers' mental models were evolving as they progressed toward adaptive expertise. In these figures, we again see the slopes of control teachers as being relatively flat while there is evidence of growth among treated classrooms.

To consolidate the six piecewise analyses just described, we examined one final model in which the outcome was the ratio of the proportion of aspirational versus transitional codes in a measure that simultaneously includes launches, rejoinders, and student contributions. Statistically significant differences were observed during the coaching interval (see Table 7, Model 3; $\beta_{21} = 1.253$, $p = .002$, ES = 2.45), meaning the overall pattern across these measures is aligned with our theory of change. As coaching proceeded, teachers began to incorporate higher proportions of dialogic talk moves, and students provided a greater proportion of strong contributions. We interpret students' growing substantive participation in discussion as evidence that teachers learned how to more skillfully employ talk moves to elicit and build on student thinking (i.e., they developed adaptive expertise).

Correspondingly, the proportion of transitional moves and weak student contributions declined. The addition of Online Content-Focused-Coaching, by itself, accounted for 52% of the variance in the growth slope during the coaching interval. Finally, with the addition of the treatment effect, there is also a main effect of initial text discussion quality on the linear growth of this ratio during the coaching phase (see Table 7, Model 3; $\beta_{22} = -0.443$, $p = .011$). All of this suggests that the coaching was critical in aiding teachers to more successfully elicit and facilitate student contributions. Furthermore, as illustrated in Figure 4, treated teachers who began with lower text discussion quality had even higher linear growth during the coaching phase.

Illustrative Example

Appendix G provides examples of classroom discussions for a single teacher before the workshop (baseline), immediately following the workshop (postworkshop), and again at the end of the year after four cycles of coaching (postcoaching). At baseline, the class discussion of *Nightmares!* by Jason Segel and Kirsten Miller (2014) very much followed a traditional (IRE) pattern. Although the teacher initiated with a partially open-ended question ("So, what is the main problem in the story?"), all of the follow-ups to students' responses either evaluated the response ("OK, I like that one") or narrowed toward a literal answer ("Are they the same person or different people, Charlotte and the

TABLE 6
Sensitivity of Treatment Effects on Text Discussion Quality to Model Selection

	Rigorous text and Questioning the Author implementation composite		Teacher facilitation moves composite		Rigor of student contributions composite	
Model	Coefficient	Standard error	Coefficient	Standard error	Coefficient	Standard error
Ordinary least squares (n = 31)	0.84**	0.25	1.08***	0.15	1.95***	0.23
Ordinary least squares with propensity score (n = 27)	0.86**	0.25	0.96***	0.13	2.01***	0.23
Ordinary least squares with propensity score strata (n = 27)	0.78**	0.26	0.98***	0.13	1.90***	0.22
Differences-in-differences (n = 31)	0.73†	0.38	0.61**	0.22	1.49**	0.46

**p < .01. ***p < .001. †p < .1.

witch?"), thus not exhibiting joint sense making of the text events. Instead, information was funneled through the teacher, who determined the correctness of students' responses. Thus, we see limited opportunities for students to drive the direction of the discussion or elaborate on their responses. When a student responded with an intriguing answer, "I think a little bit the same people," the teacher did not press for explanation, and the idea remained undeveloped. Similarly, when a student began to describe his perspective of the main problem faced by a character, the teacher seemed to direct the student's answer toward the problem that the character's mother had died, which may or may not have been the problem the student had been starting to talk about. In all, then, students had very little opportunity to explore novel ideas or interpretations, build on one another's thinking, or participate in shaping the intellectual trajectory of the discussion. The conversation was teacher centered, with little evidence of the teacher facilitating students' reasoning or elaborated responses.

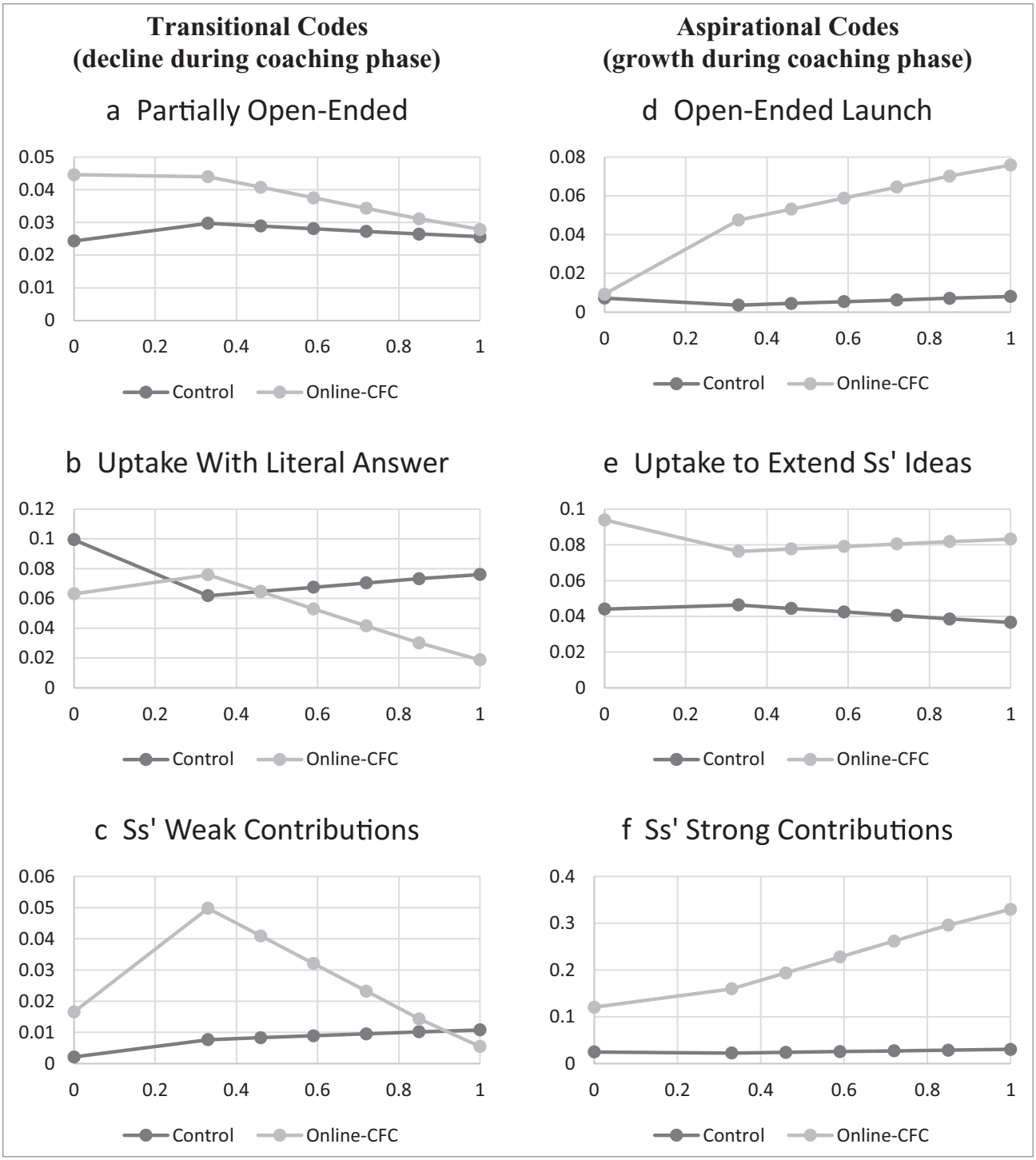
Immediately following the workshop, we see elements of Questioning the Author and Accountable Talk moves in the discussion (see Appendix G). In this lesson, aligned with Questioning the Author, the teacher read the short story *A Game of Catch* by Richard Wilbur (1994), one of the texts from the workshop, aloud to students, stopping to pose open-ended questions along the way (e.g., "What happened now?"). In contrast to the discussion at the beginning of the year (at baseline), the teacher now invited students to respond to one another (e.g., "Can you add on to that, Crystal, what Javier told us?"). As in her earlier discussion (baseline), however, the teacher continued to hold tight to the intellectual control of the discussion. For example, she quickly followed her open-ended question with a narrow question directing students to attend to a specific part of the text ("So, how are Monk and Glennie taking all of this?") and continued the practice of evaluating students' contributions. In all, then, despite the presence of new talk moves, we see limited evidence of the teacher facilitating students to deepen and extend one another's ideas and also limited evidence of the teacher pressing for student reasoning or justification for their ideas.

After four cycles of coaching, however, the discussion displayed elements more emblematic of dialogism. The teacher continued to implement Questioning the Author techniques, reading texts aloud and stopping to pose open-ended questions along the way. As shown in Appendix G, in their text discussion about the novel *True (…Sort Of)* by Katherine Hannigan (2011), students talked with one another to figure out the story (e.g., the motivation of the story character Novello). The teacher invited students to extend one another's idea to further an exploration of the idea (e.g., "How do you feel about what Mario said? Do you agree or disagree?"). The teacher pressed students to fully express their ideas by asking multiple students simply, "Why?" or "Why do you think that?" and students provided longer responses that explained their reasoning. In one instance, over several turns, the teacher facilitated a student's response to assist her in developing her idea around why Novello is teasing Delly, the main character in the story. The end result was a fully expressed and supported idea ("I don't think he's trying to get Delly to tease him.…Because in a couple chapters ago, almost the beginning of the book, it said he loved Delly"). In this short exchange, we see evidence of multiple students airing their reasoning and providing text evidence for their sense making around ideas in the text.

Changes in Teachers' Self-Reported Beliefs and Practices

Commensurate with changes in video-recorded text discussion quality, we noted changes in treated teachers' endorsed

FIGURE 3
Visual Representations of Patterns of Change for Transitional Versus Aspirational Codes From Piecewise Linear Growth Models



Note. Online-CFC = Online Content-Focused Coaching; Ss' = students'. Models adjust for teachers' content knowledge, as well as classroom characteristics: students' prior reading achievement and percentage of LatinX students in the classroom. Growth estimates during the workshop and coaching intervals adjust for initial discussion quality at baseline.

beliefs and practices with respect to discussion quality. Specifically, aligned with our theory of change, we found a statistically significant treatment contrast from baseline to spring, with declines for treated teachers in their endorsement of teacher-led discussion following IRE patterns (ES = 1.61, $p < .001$; see Table D1 for items and Table D2 for

TABLE 7
Ratio of Proportion of Dialogic Discussion Relative to Proportion of Transitional Moves

	Model 1[a]		Model 2[a]		Model 3[a]	
Covariate	Coefficient	Standard error	Coefficient	Standard error	Coefficient	Standard error
Intercept	0.992***	0.068	0.994***	0.057	0.974***	0.068
Treatment, β_{01}					0.081	0.168
Baseline text discussion quality, β_{02}			0.254***	0.068	0.239**	0.074
Linear growth estimate—workshop	0.095	0.087	0.086	0.088	−0.006	0.112
Treatment, β_{11}					0.237	0.260
Baseline text discussion quality, β_{12}			−0.030	0.108	−0.077	0.119
Linear growth estimate—coaching	0.139	0.138	0.155	0.139	−0.098	0.135
Treatment, β_{21}					1.253**	0.354
Baseline text discussion quality, β_{22}			−0.168	0.175	−0.443*	0.157
Variance components						
Intercept (percentage of reduction from the previous model)	0.028		0.007 (75%)		0.010 (0%)	
Linear growth—workshop (percentage of reduction from the previous model)	0.026		0.070 (0%)		0.087 (0%)	
Linear growth—coaching (percentage of reduction from the previous model)	0.337		0.389 (0%)		0.190 (**52%**)	
Between timepoints	0.107		0.088 (18%)		0.082 (7%)	

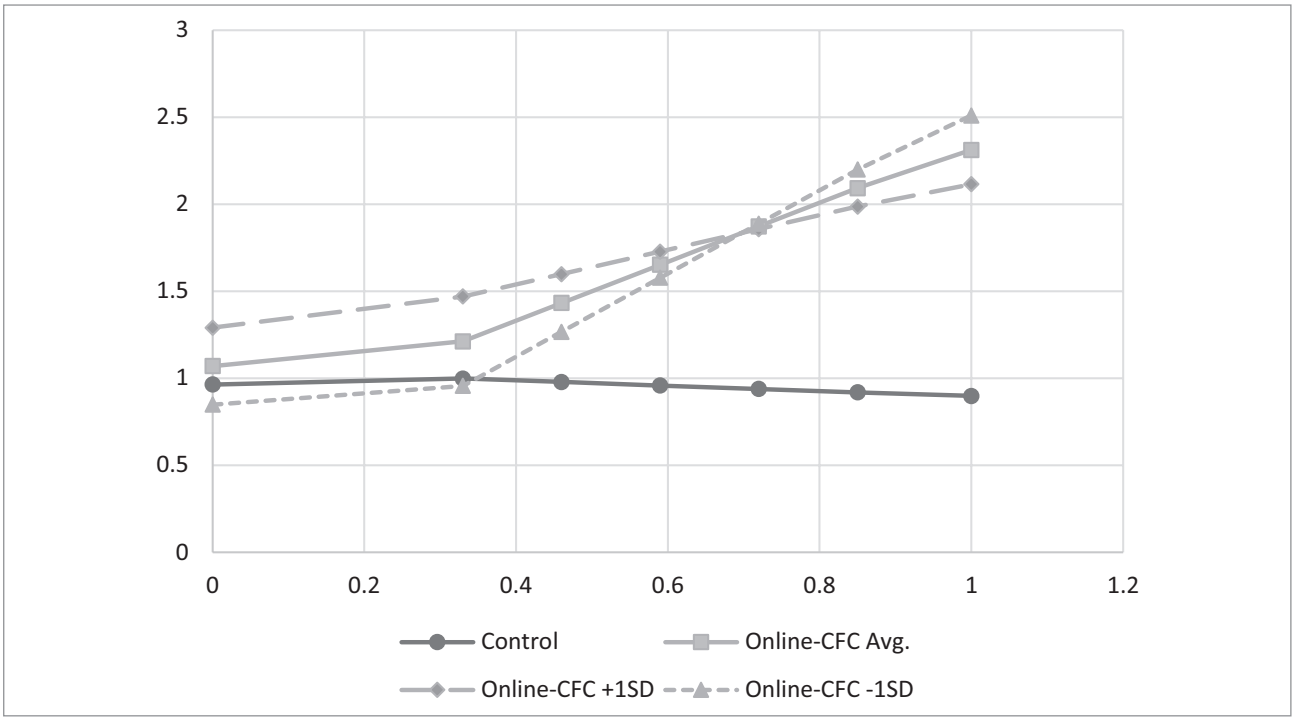
[a]Model 1 accounts for classroom-level covariates such as prior year reading achievement, percentage economically disadvantaged, average days absent, percentage minority, grade level, percentage gifted and talented, and percentage limited English proficient. Model 2 contains the same covariates as the previous model with an adjustment for starting text discussion quality. Model 3 contains the same covariates as the previous model but adds assignment condition where treatment = 1, and control = 0.
*p < .05. **p < .01. ***p < .001.

graphical depictions of differences-in-differences estimates), basic-skills comprehension instruction (ES = 1.18, p = .04), and a focus on independent reading of readily accessible texts (ES = 1.52, p < .001). Simultaneously, also aligned with a developmental hypothesis, we see evidence that teachers began to incorporate elements of dialogism in their practice. Treated teachers had higher growth relative to control teachers in their endorsement that complex texts should form the basis for whole-class discussions (ES = 1.18, p = .04; see Tables D1 and D2), and marginally statistically significant growth in their implementation of skills aligned with Questioning the Author (ES = 1.00, p = .06) and in their reports of students providing meaningful contributions to advance text discussions (ES = 0.89, p = .10). Most of the change in teachers' reports occurred during the coaching phase of Online Content-Focused Coaching. Given that control teachers' reports were consistent from baseline to spring, we interpret the changes in treated teachers (both declines and gains) as evidence that as teachers engaged in sustained reflection during coaching, their mental models (and discussion practice) shifted to be more aligned with dialogic discussions.

What Is the Relation Between Growth in Text Discussion Quality and Students' Reading Achievement?

As a final step to investigate our theory of action, we examined whether changes in teaching correlated with classroom value-added scores estimated across all available classrooms. Both bivariate and partial correlations[13] (r = .54 and .73, respectively) demonstrate a statistically

FIGURE 4
Visual Representations of the Piecewise Linear Growth Model for Ratio of Proportion of Aspirational Versus Transitional Moves



Note. Avg. = average; Online-CFC = Online Content-Focused Coaching; Online-CFC +1SD = the interaction between the effect of Online-CFC and starting one standard deviation higher on text discussion quality; Online-CFC −1SD = the interaction between the effect of Online-CFC and starting one standard deviation lower on text discussion quality. Models adjust for teachers' content knowledge, as well as classroom characteristics: students' prior reading achievement and percentage of Latinx students in the classroom. Intercept and growth estimates adjust for initial discussion quality at baseline.

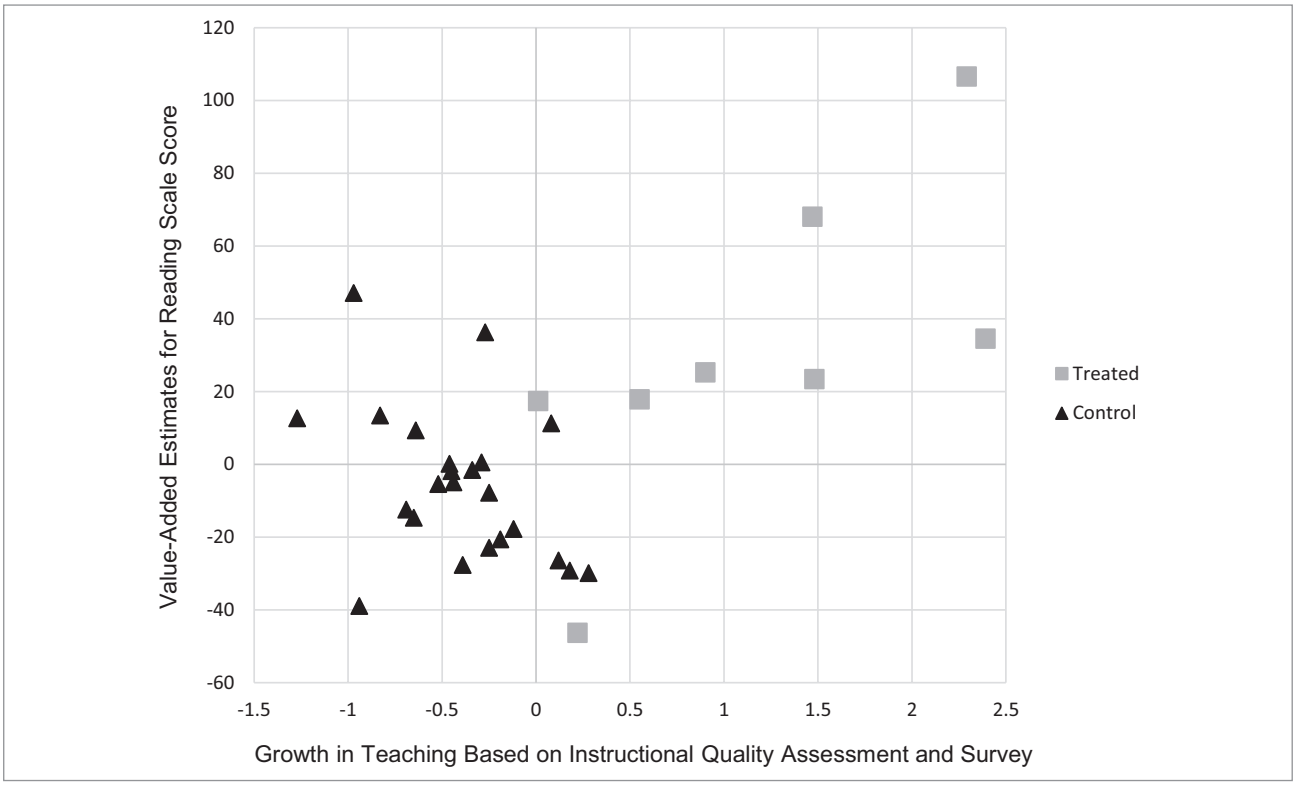
significant relation, suggesting that improvements in teaching were associated with student achievement value-added estimates. The scatterplot in Figure 5 demonstrates how these correlations may have been influenced by the treatment because seven of the eight treated teachers account for the highest estimates of teaching growth paired with the highest estimates for classroom value-added scores. Figure 5 also suggests that within the treatment group, there is a strong association between teaching growth and the magnitude of the classroom value-added scores on reading achievement. In other words, aligned with our hypothesis (theory of change), substantive increases in teachers' skillful use of talk moves (i.e., developing adaptive expertise) was highly related to growth in students' reading comprehension skills.

Discussion

Hypothesis Tests of Online Content-Focused Coaching in the Context of Coaching Studies

Our study was designed as a small-scale efficacy trial for Online Content-Focused Coaching. Results from our study demonstrate a relation between teaching improvement and student achievement value-added estimates. Looking at our results in the context of other research on instructional coaching, we note that the main effects for Online Content-Focused Coaching on text discussion quality (ES > 1.12) are nearly twice as large as those reported in a meta-analysis for average effects of coaching on instruction for programs with fewer than 50 teachers (ES = 0.63; Kraft, Blazar, & Hogan, 2018). The main effect for students of treated teachers on those students' reading achievement (ES = 0.24) on STAAR are comparable to effects of coaching on achievement for small programs in the same meta-analysis (ES = 0.28; Kraft et al., 2018). Furthermore, whereas Kraft et al. (2018) demonstrated that the size of the program is one potential factor influencing the magnitude of the effect size, another is whether the achievement test used to demonstrate the effect is project developed versus a (state) standardized assessment. Blank and de las Alas (2009) found that studies using (project-developed) measures aligned with the focus of the professional development had a mean effect size of 0.32 as compared with a mean effect size of 0.10 for studies using statewide assessments as an outcome. Because smaller programs are more likely to have the capability to administer assessments more aligned with

FIGURE 5
Association Between Changes in Teaching and Changes in Students' Comprehension Scores by Condition Within the Analytic Sample



their intervention, we think this might also be a factor in smaller studies demonstrating a greater magnitude effect size. We find it notable that our effects on achievement were found on STAAR and, thus, reflect constructs measuring reading skills that are more likely to generalize across populations.

Limitations of these findings are due largely to the sample size and differential attrition. Although teachers in the treated group taught in similar contexts and were quite similar to teachers in the control group, we have limited ability to generalize the treatment effect to all potential teachers because we do not know why the eight teachers completing the treatment continued the study after random selection and why other teachers who had been selected for treatment did not continue. We know that 32% of treated teachers who were recruited in the spring did not enlist in baseline activities in the fall. Therefore, some teacher turnover occurred during the summer before the opportunity to participate in treatment even began. We also know that the time of year had an influence on attrition rates, as the attrition rate from the workshop was greater for the treated group in the fall (47%) than it was for the control group in the spring (9%). Replication studies will seek to implement strategies to reduce attrition and develop findings that work toward population-level generalizations.

We note, however, that we conducted a thorough analysis of missing data in response to differential attrition in our design and concluded that we could establish moderate baseline equivalence between treated and control teachers on prior achievement and on most demographic measures. In addition, in all of our analyses, including those in which baseline equivalence was not established, we included covariates to adjust for, among other things, students' prior achievement, students' background, and teachers' baseline measures of text discussion quality.

Theorizing and Observing a Process of Change in Dialogic Text Discussion Quality

We have taken seriously the charge for education researchers to develop and test theories around how teachers' learning in practice-based professional development influences teaching and student outcomes (see, e.g., Kennedy, 2016), in addition to providing traditional evidence for the efficacy of our intervention as described earlier. Such theorizing, or ability to explain how and why interventions influence change in desired outcomes, is central to establishing consequential validity (Moss, 2016) and, more practically, for designing interventions that significantly

increase discussion quality, both with respect to teacher facilitation moves and students' expression of higher level thinking. This latter point is important given the critical role that dialogic discussions play in student learning (see, e.g., Wilkinson et al., 2015) and the known difficulty that teachers have in generating and facilitating students' disciplinary reasoning in discussions (Franke et al., 2009).

Our work draws on theories from cognitive scientists about modern, dynamic models of conceptual change, that is, a process of change in which alternative actions to conflicting beliefs, aims, or values are considered, decided on, and justified. This process inevitably leads to a revision of existing mental models, thereby building teachers' capacity toward adaptive expertise in discussion facilitation. In the context of our intervention, we view both the workshop and the coaching components of Online Content-Focused Coaching as essential for growth in adaptive expertise. The workshop set the foundation for teacher–coach conversations that led to teaching growth. Teacher learning was facilitated by establishing common aims for text discussions, by providing a common framework and vocabulary that grounded the teacher–coach discussions, and by scaffolds for planning text discussions that improved teachers' willingness to attempt new talk moves during initial attempts to implement Questioning the Author and Accountable Talk (the first identified stage of improvement).

This last point was critical for the video-based coaching because these attempts to implement new moves were constructive for teacher and coach reflections on whether they succeeded with students initially or not. The importance of these initial attempts to elicit student thinking lay in the opportunity for the teacher, with the aid of a coach, to develop an inquiry stance for reflexively considering these talk moves in relation to student responses vis-à-vis their student learning goals. A critical element for this work of coaching was the coach helping teachers recognize instances of cognitive conflict so teachers were positioned to reflect on them to suggest and consider alternative talk moves (with the coach providing scaffolded support). The resolution and explanation of a well-reasoned alternative produces an atomistic element: a causal attribution among a teacher's problem space (e.g., "My students stare at me with blank looks after an open-ended question"), a talk move (e.g., "I wait but also continue to invite participation"), and why this would result in desired student contribution(s). We view the development of adaptive expertise as an accumulation of these atomistic associations. Furthermore, subsequent generalizations of these associations about when and why a teacher would employ similar talk moves are viewed as enabling teachers to enact real-time facilitation moves during discussions.

We note that our theorizing about the process of change for Online Content-Focused Coaching is supported by our prior analyses with different cohorts of teachers participating in the intervention (Matsumura et al., 2019). In that work, two-rate piecewise growth models of change in talk moves helped us infer the process of development, which was aligned with the timing of the workshop and coaching. We assumed that improvements from the workshop would generate some open-ended questions from teachers and also partially open-ended questions during teachers' initial attempts to implement aspects of Questioning the Author and Accountable Talk moves. Likewise, we presumed that facilitation moves would follow a similar pattern because teachers were just beginning to implement facilitation moves consistent with their evolving mental models for text discussions. Indeed, this is what we observed in the current study as well. Teachers demonstrated growth in transitional moves, such as partially open-ended talk moves immediately after the workshop. In the next phase of Online Content-Focused Coaching, during successive cycles of coaching, we observed a decline in transitional moves and an increase in aspirational moves (open-ended talk moves). In particular, strong student contributions increased during this phase of the intervention. This change coincided with an increase in the rate of aspirational moves and a decrease in the rate of transitional moves. Thus, during the coaching, we observed changes in teachers' facilitation of text discussions, which we attribute to growth in adaptive expertise.

Measuring Change in Dialogic Text Discussion Quality

In addition to interrogating a theory of change toward adaptive expertise, our research also sets out a measurement strategy that could be useful for gauging the development of dialogic instruction. As noted earlier, knowledge of both a developmental process and a measurement approach is crucial for translating research into practical improvements (i.e., consequential validity). Critical to our inferences was an exhaustive coding of discussions, particularly student contributions, which were important for making formative[14] judgments about changes over time in our treated classrooms relative to the control classrooms in our study. As students' weak contributions increased and then declined during coaching (see Figure 3c), we also observed a steep increase in students' strong contributions during coaching (see Figure 3f), thus providing empirical support that teachers' in-the-moment facilitation of students' responses had shifted.

In the context of an intervention designed with the objective of promoting dialogic text discussions in which students are positioned to do the thinking, we saw evidence of this in our coding of student contributions. Thus, the observable student contributions were the cornerstone demonstrating a tangible change in text discussions, that is, direct evidence of student thinking and reasoning. In addition, these moments when students demonstrated their

reasoning were also critical to further advance teachers' mental models toward adaptive expertise, as they were fodder for contrastive cases in which teachers could analyze moves that led to students' reasoning during discussion. Finally, for researchers, moments of student reasoning can be used to understand the relation between teaching moves and student reasoning or, as was demonstrated here, identify a process of development toward adaptive expertise.

Conclusion

We presented findings from an intervention study that demonstrated effects on text discussion quality and student learning, and we also made connections between the magnitude of change in text discussion practices and growth in student achievement. Although teachers varied in their rates of change in text discussion quality, we demonstrated an evolution of teachers' text discussion practices. Through exhaustive coding of repeated video-recorded discussions, we were able to identify different stages of improvement. In our small randomized sample, this multistage process roughly corresponded to what we see as the strengths of the workshop and coaching components of Online Content-Focused Coaching. We propose a theory for the process of change aligned with our observed changes in teaching talk moves and students' strong text-based contributions during different phases of the intervention. In our view, having teachers consider and explain cause–effect relations for how teaching moves influence student contributions in the context of their existing beliefs, aims, and values is essential for conceptual change in teaching, yielding movement toward teachers' adaptive expertise in facilitating discussions.

NOTES

This work was supported by a grant (R305A140394) from the Institute of Education Sciences, U.S. Department of Education. The views expressed herein are those of the authors and have not been reviewed or approved by the Institute of Education Sciences or the U.S. Department of Education. The work was also supported by a Teachers-as-Learners grant from the James S. McDonnell Foundation.

[1]We note the similarities between the terms *schemata* and *mental models*. Throughout the text, we refer to such dynamic knowledge structures in the teachers' mind as mental models because of their modern connection to simulation, reasoning, and decision making, which all have consequence for teachers' discussion facilitation.

[2]We note that Figure 1 represents an oversimplification of the learning process in two ways. First, learning is both dynamic and fluid. Prior conceptions are not replaced completely, but rather learners adopt more or less dominant conceptions that are linked to changing contexts (Nadelson, Heddy, Jones, Taasoobshirazi, & Johnson, 2018). Second, whereas we present a view of conceptual change for teaching as staged, the rate that teachers progress through the stages is more individualized than is represented here.

[3]We define the term *reflexively* as used in this context to mean having teachers build their own analytic reasoning (an inquiry stance) specifically vis-à-vis connections they make between their own practice and consequently the student thinking elicited by their talk moves in their classroom text discussions. Thus, the examination of cause–effect associations between teaching moves and student thinking might be facilitated by the coach, with the intended goal of building the teachers' capacity to routinely engage in their own reflexive practice subsequently.

[4]We examined *t*-tests across the 27 participating schools on all of the demographic characteristics that we had access to from the year prior to randomization. We also examined *t*-tests across the 62 potential participating teachers and found balance on the same set of demographic characteristics at the teacher level, as well.

[5]In addition, although not statistically significantly different, the standardized mean difference (Hedges' *g*) for the other two IQA composite scores are both positive and trend toward a difference ($|g| = 0.18$ for the choosing rigorous texts and implementing elements of Questioning the Author composite, and $|g| = 0.51$ for the rigor of student contributions composite).

[6]Our data would be considered fully balanced panel data because no teachers were missing at baseline, nor were they missing an end-of-year measure, resulting in no missing teachers due to listwise deletion.

[7]We consider weak student contributions a transitional move because one goal for dialogic discussions is to increase the number of student contributions. Two things are likely necessary to enable strong student contributions: authentic open-ended questions and a teacher's strategic employment of facilitation moves. Consistent with our theory, we think many teachers will begin coaching with a desire to increase student contributions, but their students will struggle to provide strong student contributions without support, and teachers' will lack the adaptive expertise to facilitate strong contributions.

[8]We compared mean differences for the treated and control analytic sample prior to intervention on each of the six constructs. The means were not statistically significantly different between groups (αssss = .05), but Hedges' *g* was between 0.25 and 0.30 for three of the six constructs (see Table B2 in Appendix B).

[9]Although these models omit these four teachers, the findings are similar to the initial models that we ran comparing the treated teachers with all other teachers ($n = 48$ classrooms) for whom we had achievement data in the district, as well as a completed baseline survey, so we were able to make linear adjustments for statistical control variables.

[10]These estimates may be slightly optimistic or may indicate a dosage effect of the coaching. We note that the average of the last coached videos was 3.17 for the rigorous student contributions composite. Although this indicates that not all teachers attained a score of 4 on their last coached video, the discrepancy between growth estimates being near 4 and the previously reported mean of 3.17 is due to the fact that teachers with more coached sessions scored, on average, close to 4 on this measure. In other words, the hierarchical linear growth model estimates the group trajectory relative to time for all teachers for all of their coached videos, revealing the underlying growth trajectory. We would trust growth estimates more than the snapshot mean of last coached video, because they represent the best expression of interindividual change, if we felt that the assumption held that data were missing completely at random. However, we do not have evidence to confirm or disconfirm that assumption.

[11]We examined a series of latent variable regressions adjusting for the starting point of teachers to understand whether initial discussion quality confounded our measurement of growth, and we examined prediction models with covariates from the pre-survey ($n = 27$ teachers), including measures of teachers' background and prior knowledge.

[12]In the face-to-face Content-Focused Coaching work, effect size estimates on the IQA were slightly smaller (ES = .70), but the studies are not strictly comparable because they reflect a difference in scale (approximately 170 teachers were involved in the face-to-face

Content-Focused Coaching study) as much as a difference in delivery method of the treatment.

[13]Partial correlations were adjusted for teacher characteristics such as their pedagogical content knowledge for teaching score (see, e.g., Phelps & Gitomer, 2012) and their years of experience, and an additional control for school prior achievement in writing.

[14]Although we report changes in student achievement on STAAR as evidence of change in students' reading skills, we also assert that direct observation of student learning inheres in text discussions themselves. Indeed, formative assessment within text discussions (i.e., teachers thinking reflexively about how students' contributions to text discussions demonstrate their reasoning) is aligned with sociocultural learning theories. If research were also aligned to measures of students' reasoning in discussions, then research findings could be more aligned to assessment and to modern learning theories. Such formative assessment would connect with dialogic learning goals while also promoting and supporting the instructional change we desire (see, e.g., Shepard, Penuel, & Pellegrino, 2018).

[15]Teachers reported leaving the study for various non-study-related reasons. Primary among these reasons was a change in position either to a different grade level or for a promotion out of the classroom ($n = 5$), self or family health-related issues ($n = 5$), or other personal reasons.

[16]The last video was collected just prior to the state standardized testing period in April.

[17]In addition to exploring potential answers to this question with our baseline data, we attempted to account for differential attrition in our analytic methods.

[18]We list several of these (e.g., change of position, no longer teaching English language arts, self and family medical issues), but it is impossible to verify those reports.

[19]Definitions for most of the teacher moves are taken from the Analyzing Teaching Moves codebook (Correnti et al., 2015).

[20]However, if the teacher asks more than one question of the same type in succession, code it only once. This is because, generally, they are trying out different phrasings of the question (e.g., "Why do you think that?"; "What is there in the text?"; "What makes you say that?").

[21]In our analyses, we centered time so the intercept was the estimated IQA dimension score for teachers at the beginning of the study.

[22]The indicator for participation in Online Content-Focused Coaching was entered in the model uncentered, meaning the estimates for the intercept are for control teachers only.

REFERENCES

Alexander, P.A. (2003). The development of expertise: The journey from acclimation to proficiency. *Educational Researcher*, *32*(8), 10–14. https://doi.org/10.3102/0013189X032008010

Alexander, P.A. (2004). A model of domain learning: Reinterpreting expertise as a multidimensional, multistage process. In D.Y. Dai & R.J. Sternberg (Eds.), *Motivation, emotion, and cognition: Integrative perspectives on intellectual functioning and development* (pp. 273–298). Mahwah, NJ: Erlbaum.

Alexander, R. (2006). *Towards dialogic teaching: Rethinking classroom talk* (4th ed.). York, UK: Dialogos.

Anderson, R.C. (1984). Role of the reader's schema in comprehension, learning, and memory. In R.C. Anderson, J. Osborn, & R.J. Tierney (Eds.), *Learning to read in American schools: Basal readers and content texts* (pp. 243–257). Hillsdale, NJ: Erlbaum.

Applebee, A.N., Langer, J.A., Nystrand, M., & Gamoran, A. (2003). Discussion-based approaches to developing understanding: Classroom instruction and student performance in middle and high school English. *American Educational Research Journal*, *40*(3), 685–730. https://doi.org/10.3102/00028312040003685

Beck, I.L., & McKeown, M.G. (2001). Text talk: Capturing the benefits of read-aloud experiences for young children. *The Reading Teacher*, *55*(1), 10–20.

Beck, I.L., & McKeown, M.G. (2006). *Improving comprehension with Questioning the Author: A fresh and expanded view of a powerful approach*. New York, NY: Scholastic.

Blank, R.K., & de las Alas, N. (2009). *Effects of teacher professional development on gains in student achievement: How meta analysis provides scientific evidence useful to education leaders*. Washington, DC: Council of Chief State School Officers.

Boyd, M.P., & Markarian, W.C. (2015). Dialogic teaching and dialogic stance: Moving beyond interactional form. *Research in the Teaching of English*, *49*(3), 272–296.

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

Chi, M.T.H. (2008). Three types of conceptual change: Belief revision, mental model transformation, and categorical shift. In S. Vosniadou (Ed.), *International handbook of research on conceptual change* (pp. 61–82). New York, NY: Routledge.

Chinn, C.A., Anderson, R.C., & Waggoner, M.A. (2001). Patterns of discourse in two kinds of literature discussion. *Reading Research Quarterly*, *36*(4), 378–411. https://doi.org/10.1598/RRQ.36.4.3

Collins, A. (2006). Cognitive apprenticeship. In R.K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (pp. 47–60). New York, NY: Cambridge University Press.

Collins, A., Brown, J.S., & Holum, A. (1991). Cognitive apprenticeship: Making thinking visible. *American Educator*, *6*(11), 38–46.

Correnti, R., Stein, M.K., Smith, M.S., Scherrer, J., McKeown, M., Greeno, J., & Ashley, K. (2015). Improving teaching at scale: Design for the scientific measurement and learning of discourse practice. In L.B. Resnick, C.S.C. Asterhan, & S.N. Clarke (Eds.), *Socializing intelligence through academic talk and dialogue* (pp. 315–332). Washington, DC: American Educational Research Association.

Franke, M.L., Webb, N.M., Chan, A.G., Ing, M., Freund, D., & Battey, D. (2009). Teacher questioning to elicit students' mathematical thinking in elementary school classrooms. *Journal of Teacher Education*, *60*(4), 380–392. https://doi.org/10.1177/0022487109339906

Gafoor, K.A., & Akhilesh, P.T. (2010). Strategies for facilitating conceptual change in school physics. *Innovations and Researches in Education*, *3*(1), 34–42.

Goldman, S.R., Snow, C., & Vaughn, S. (2016). Common themes in teaching reading for understanding: Lessons from three projects. *Journal of Adolescent & Adult Literacy*, *60*(3), 255–264. https://doi.org/10.1002/jaal.586

Greeno, J.G., Collins, A.M., & Resnick, L.B. (1996). Cognition and learning. In D.C. Berliner & R.C. Calfee (Eds.), *Handbook of educational psychology* (pp. 15–46). New York, NY: Macmillan.

Gregoire, M. (2003). Is it a challenge or a threat? A dual-process model of teachers' cognition and appraisal processes during conceptual change. *Educational Psychology Review*, *15*(2), 147–179. https://doi.org/10.1023/A:1023477131081

Grossman, P.L., Smagorinsky, P., & Valencia, S. (1999). Appropriating tools for teaching English: A theoretical framework for research on learning to teach. *American Journal of Education*, *108*(1), 1–29. https://doi.org/10.1086/444230

Hatano, G., & Inagaki, K. (1986). Two courses of expertise. In H. Stevenson, H. Azuma, & K. Hakuta (Eds.), *Child development and education in Japan* (pp. 262–272). New York, NY: Freeman.

Heckman, J.J. (1979). Sample selection bias as specification error. *Econometrica*, *47*(1), 153–161. https://doi.org/10.2307/1912352

Ivers, N.M., Halperin, I.J., Barnsley, J., Grimshaw, J.M., Shah, B.R., Tu, K., … Zwarenstein, M. (2012). Allocation techniques for balance at baseline in cluster randomized trials: A methodological review. *Trials*, *13*, article 120. https://doi.org/10.1186/1745-6215-13-120

Johnson-Laird, P.N. (1980). Mental models in cognitive science. *Cognitive Science*, *4*(1), 71–115. https://doi.org/10.1207/s15516709cog0401_4

Kennedy, M.M. (2016). How does professional development improve teaching? *Review of Educational Research*, *86*(4), 945–980. https://doi.org/10.3102/0034654315626800

Kintsch, W., & van Dijk, T.A. (1978). Toward a model of text comprehension and production. *Psychological Review*, *85*(5), 363–394. https://doi.org/10.1037/0033-295X.85.5.363

Koo, T.K., & Li, M.Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Kraft, M.A., Blazar, D., & Hogan, D. (2018). The effect of teacher coaching on instruction and achievement: A meta-analysis of the causal evidence. *Review of Educational Research*, *88*(4), 547–588. https://doi.org/10.3102/0034654318759268

Kucan, L. (2009). Engaging teachers in investigating their teaching as a linguistic enterprise: The case of comprehension instruction in the context of discussion. *Reading Psychology*, *30*(1), 51–87. https://doi.org/10.1080/02702710802274770

Langer, J.A. (1995). *Envisioning literature: Literary understanding and literature instruction*. Newark, DE: International Reading Association; New York, NY: Teachers College Press.

Lefstein, A., Snell, J., & Israeli, M. (2015). From moves to sequences: Expanding the unit of analysis in the study of classroom discourse. *British Educational Research Journal*, *41*(5), 866–885. https://doi.org/10.1002/berj.3164

Lohr, S., Schochet, P.Z., & Sanders, E. (2014). *Partially nested randomized controlled trials in education research: A guide to design and analysis* (NCER 2014-2000). Washington, DC: National Center for Education Research, Institute of Education Sciences, U.S. Department of Education.

Matsumura, L.C., Correnti, R., Walsh, M., Bickel, D.D., & Zook-Howell, D. (2019). Online Content-Focused Coaching to improve classroom discussion quality. *Technology, Pedagogy and Education*, *28*(2), 191–215. https://doi.org/10.1080/1475939X.2019.1577748

Matsumura, L.C., Garnier, H.E., Correnti, R., Bickel, D.D., & Junker, B. (2010). Investigating the effectiveness of a comprehensive literacy-coaching program in schools with high teacher mobility. *The Elementary School Journal*, *111*(1), 35–62. https://doi.org/10.1086/653469

Matsumura, L.C., Garnier, H.E., Slater, S.C., & Boston, M.D. (2008). Toward measuring instructional interactions "at-scale". *Educational Assessment*, *13*(4), 267–300. https://doi.org/10.1080/10627190802602541

Matsumura, L.C., Garnier, H.E., & Spybrook, J. (2012). The effect of Content-Focused Coaching on the quality of classroom text discussions. *Journal of Teacher Education*, *63*(3), 214–228. https://doi.org/10.1177/0022487111434985

Matsumura, L.C., Garnier, H.E., & Spybrook, J. (2013). Literacy coaching to improve student reading achievement: A multi-level mediation model. *Learning and Instruction*, *25*, 35–48. https://doi.org/10.1016/j.learninstruc.2012.11.001

McMillan, J.H. (2007). Randomized field trials and internal validity: Not so fast my friend. *Practical Assessment, Research & Evaluation*, *12*(1), article 15. https://doi.org/10.7275/3vh7-m792.

McNeill, K.L., & Pimentel, D.S. (2010). Scientific discourse in three urban classrooms: The role of the teacher in engaging high school students in argumentation. *Science Education*, *94*(2), 203–229. https://doi.org/10.1002/sce.20364

Mehan, H. (1979). *Learning lessons: Social organization in the classroom*. Cambridge, MA: Harvard University Press.

Michaels, S., O'Connor, C., & Resnick, L.B. (2008). Deliberative discourse idealized and realized: Accountable talk in the classroom and in civic life. *Studies in Philosophy and Education*, *27*, 283–297. https://doi.org/10.1007/s11217-007-9071-1

Moss, P.A. (2016). Shifting the focus of validity for test use. *Assessment in Education: Principles, Policy & Practice*, *23*(2), 236–251. https://doi.org/10.1080/0969594X.2015.1072085

Munter, C., & Correnti, R. (2017). Examining relations between mathematics teachers' instructional vision and knowledge and change in practice. *American Journal of Education*, *123*(2), 171–202. https://doi.org/10.1086/689928

Murphy, P.K., Greene, J.A., Firetto, C.M., Hendrick, B.D., Li, M., Montalbano, C., & Wei, L. (2018). Quality talk: Promoting students' discourse to promote high-level comprehension. *American Educational Research Journal*, *55*(5), 1113–1160. https://doi.org/10.3102/0002831218771303

Murphy, P.K., Wilkinson, I.A.G., Soter, A.O., Hennessey, M.N., & Alexander, J.F. (2009). Examining the effects of classroom discussion on students' comprehension of text: A meta-analysis. *Journal of Educational Psychology*, *101*(3), 740–764. https://doi.org/10.1037/a0015576

Nadelson, L.S., Heddy, B.C., Jones, S., Taasoobshirazi, G., & Johnson, M. (2018). Conceptual change in science teaching and learning: Introducing the dynamic model of conceptual change. *International Journal of Educational Psychology*, *7*(2), 151–195. https://doi.org/10.17583/ijep.2018.3349

National Center for Education Statistics. (2017). *The Nation's Report Card: National Assessment of Educational Progress reading assessment*. Washington, DC: National Center for Education Research, Institute of Education Sciences, U.S. Department of Education.

Nystrand, M. (2006). Research on the role of classroom discourse as it affects reading comprehension. *Research in the Teaching of English*, *40*(4), 392–412.

Phelps, G., & Gitomer, D. (2012). *Content knowledge for teaching assessment development: Assessment of content knowledge for teaching for teachers of grades 4–6 English language arts*. Seattle, WA: Bill and Melinda Gates Foundation; Princeton, NJ: Educational Testing Service.

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

Rogosa, D. (1995). Myths and methods: "Myths about longitudinal research" plus supplemental questions. In J.M. Gottman (Ed.), *The analysis of change* (pp. 3–66). Mahwah, NJ: Erlbaum.

Sedova, K., Salamounova, Z., & Svaricek, R. (2014). Troubles with dialogic teaching. *Learning, Culture and Social Interaction*, *3*(4), 274–285. https://doi.org/10.1016/j.lcsi.2014.04.001

Sedova, K., Sedlacek, M., & Svaricek, R. (2016). Teacher professional development as a means of transforming student classroom talk. *Teaching and Teacher Education*, *57*, 14–25. https://doi.org/10.1016/j.tate.2016.03.005

Shadish, W.R., Cook, T.D., & Campbell, D.T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton-Mifflin.

Shepard, L.A., Penuel, W.R., & Pellegrino, J.W. (2018). Using learning and motivation theories to coherently link formative assessment, grading practices, and large-scale assessment. *Educational Measurement: Issues and Practice*, *37*(1), 21–34. https://doi.org/10.1111/emip.12189

Sherin, M.G., & van Es, E.A. (2009). Effects of video club participation on teachers' professional vision. *Journal of Teacher Education*, *60*(1), 20–37. https://doi.org/10.1177/0022487108328155

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

Soter, A.O., Wilkinson, I.A., Murphy, P.K., Rudge, L., Reninger, K., & Edwards, M. (2008). What the discourse tells us: Talk and indicators of high-level comprehension. *International Journal of Educational Research*, *47*(6), 372–391. https://doi.org/10.1016/j.ijer.2009.01.001

Spillane, J.P., Reiser, B.J., & Reimer, T. (2002). Policy implementation and cognition: Reframing and refocusing implementation research.

Review of Educational Research, *72*(3), 387–431. https://doi.org/10.3102/00346543072003387
Taylor, S.E., & Crocker, J. (1981). Schematic bases of social information processing. In E.T. Higgins, C.P. Herman, & M.P. Zanna (Eds.), *Social cognition* (pp. 89–134). Hillsdale, NJ: Erlbaum.
Texas Education Agency. (2015). *STAAR 2015 mean p-values and internal consistency values by reporting category and content area. In Technical digest 2014-2015*. Austin: Author. Retrieved from https://tea.texas.gov/sites/default/files/digest15-appendB-STAAR-Reliability.pdf
Thompson, J., Windschitl, M., & Braaten, M. (2013). Developing a theory of ambitious early-career teacher practice. *American Educational Research Journal*, *50*(3), 574–615. https://doi.org/10.3102/0002831213476334
Vigital. (2015). *Studiocode software 5.8.6* [Computer software]. Author.
Vygotsky, L.S. (1986). *Thought and language* (A. Kozulin, Trans.). Cambridge, MA: MIT Press. (Original work published 1934)
Wilke, R.A., & Losh, S.C. (2012). Exploring mental models of learning and instruction in teacher education. *Action in Teacher Education*, *34*(3), 221–238. https://doi.org/10.1080/01626620.2012.693241
Wilkinson, I.A.G., Murphy, P.K., & Binici, S. (2015). Dialogue-intensive pedagogies for promoting reading comprehension: What we know, what we need to know. In L.B. Resnick, C.S.C. Asterhan, & S.N. Clarke (Eds.), *Socializing intelligence through academic talk and dialogue* (pp. 37–50). Washington, DC: American Educational Research Association.

LITERATURE CITED

Hannigan, K. (2011). *True (…sort of)*. New York, NY: HarperCollins.
Segel, J., & Miller, K. (2014). *Nightmares!* New York, NY: Yearling.
Wilbur, R. (1994). *A game of catch*. San Diego, CA: Harcourt Brace.

Submitted March 18, 2019
Final revision received March 24, 2020
Accepted March 26, 2020

RICHARD CORRENTI (corresponding author) is an associate professor in the School of Education and a research scientist in the Learning Research and Development Center at the University of Pittsburgh, Pennsylvania, USA; email rcorrent@pitt.edu.

LINDSAY CLARE MATSUMURA is a professor in the School of Education and a senior scientist in the Learning Research and Development Center at the University of Pittsburgh, Pennsylvania, USA; email lclare@pitt.edu.

MARGUERITE WALSH is a doctoral candidate in the Learning Sciences and Policy Program at the University of Pittsburgh, Pennsylvania, USA; email mew138@pitt.edu.

DENA ZOOK-HOWELL is a former fellow in the Institute for Learning and a research associate in the Learning Research and Development Center at the University of Pittsburgh, Pennsylvania, USA; email dkz5@pitt.edu.

DONNA DIPRIMA BICKEL is a retired senior fellow in the Institute for Learning in the Learning Research and Development Center at the University of Pittsburgh, Pennsylvania, USA; email dbickel@pitt.edu.

BAEKSAN YU is a doctoral candidate in the Department of Administrative and Policy Studies at the University of Pittsburgh, Pennsylvania, USA; email postcentre@pitt.edu.

APPENDIX A

Framework for Effective Text Discussions

Dimension	Teacher move(s)	Student move(s)
Select a complex text with grist	• Select a text with grist/complexity that supports extended responses and meaning making in discussion.	• Show interest in the topic. • Demonstrate motivation to persist and grapple with challenging content to make sense of the text.
Segment the text	• Identify stopping points during reading that provide opportunities to unpack text difficulties. • Plan initial questions and potential follow-up questions.	• Engage in making sense along the way (i.e., during reading).
Pose questions to construct the gist	• Ask open-ended questions that require students to respond in more elaborate ways to explain an idea in the text. • Ask questions that surface students' potential misunderstandings. • Ask questions in sequence that help students construct understanding of the key ideas in the text.	• Demonstrate understanding of key ideas in the text. • Respond using own words rather than repeating the text verbatim. • Respond in longer ways that connect ideas within the text.

(continued)

Framework for Effective Text Discussions (*continued*)

Dimension	Teacher move(s)	Student move(s)
Pose cognitively demanding questions	• Ask questions that link text ideas to broader issues in the discipline or world. • Ask questions that require text interpretation and analysis.	• Form generalizations, claims, and/or arguments about the text.
Develop accountability to accurate knowledge	• Mark critical ideas expressed by students. • Press for accuracy in students' responses. • Build on students' prior knowledge.	• Demonstrate accurate knowledge of the ideas in the text. • Identify knowledge not yet available but needed to address an issue.
Develop accountability to rigorous thinking	• Challenge students' explanations. • Press students to explain their reasoning. • Invite students to expand on their thinking. • Model reasoning (i.e., think aloud). • Recapitulate ideas expressed in the discussion.	• Explain own reasoning about text-based evidence. • Test understanding of concepts. • Formulate hypotheses based on text evidence. • Challenge the quality of one another's evidence and reasoning.
Develop accountability to community	• Invite participation to ensure that all (or nearly all) students participate in the discussion. • Link students' ideas in the discussion (i.e., show how critical ideas expressed by students relate to one another). • Work to keep everyone together. • Verify and clarify students' contributions to ensure that the students are understood.	• Actively participate in the discussion. • Listen attentively to one another. • Elaborate and build on one another's ideas. • Work to clarify or expand an idea.

APPENDIX B

Differential Attrition

Our first research activity asked teachers to complete a baseline survey of their teaching, focusing primarily on teachers' beliefs and practices about text-based discussion practices. Of the 31 teachers who signed up in the summer and were assigned to treatment, only 21 completed the baseline survey and contributed a baseline video. Another six treated teachers dropped from the study for various reasons before the workshop,[15] yielding a pool of 15 potential teachers for the treatment condition. Of these 15 teachers, seven of them dropped from the study within the first couple of weeks of the workshop (attrition due to the fall workshop = 47%), and the other eight continued throughout the course of the treatment, completing the workshop and coaching. In the control condition, 29 of the 31 teachers assigned to the condition completed a baseline survey. Twenty-three teachers completed the data collection, including a baseline video along with two other videos of their text discussions at evenly spaced intervals over the year.[16] We calculated two different attrition rates for the control group. First, through the end of data collection activities prior to the workshop (including all the data for identifying main effects), 23 of the 29 teachers completed the study (attrition due to data collection = 21%). Second, when provided the chance to take the workshop, 21 of these 23 teachers completed the workshop (attrition due to the spring workshop = 9%).

Differential attrition in our study poses interesting analytic challenges because attributions of causality may depend on the selection mechanism, if any exists, into treatment when assigned to that condition. Thus, one alternative explanation for the putative cause of findings could be because participants least likely to show positive change were more likely to quit the treatment.[17] Although it is impossible to know the actual mechanism for attrition,[18] we explored our data to understand whether teachers from one of our strata were more likely to drop or attrit from the study. We found that the attrition rate differed by our original school stratification (see Table B1). In our third stratum, where schools were identified to have greater than 50% of their students as English learners, there were 13 teachers who indicated interest in participating in the study in May. However, only seven of them took the baseline survey, and

TABLE B1
Differences in Teacher Attrition by School Stratum

Stratum	Number of schools	Teachers interested in May 2016	Teachers taking the baseline survey in fall 2016	Teachers completing data collection in spring 2017 (attrition rate)
<20% English learners	7	21	21	16 (24%)
20% < English learners < 50%	12	28	22	18 (18%)
>50% English learners	8	13	7	2 (71%)

only two of the teachers (both in the control condition) completed the study data collection activities (just 15% of those originally indicating interest).

Differential attrition between strata demonstrates differences across schools in the participation rates of teachers in the study, which will influence inferences about the external validity of our findings (i.e., the findings will only generalize to the teachers teaching in contexts similar to the teachers who completed the study). The fact that both treated and control teachers from this stratum were seen dropping out and experiencing attrition at very high rates led us to further investigate (a) whether teacher participants who stayed in the treatment condition (n = 8) were different from those who left the treatment condition (n = 7), (b) whether teachers who dropped from the control condition were different from those who remained in the control condition, and (c) whether the treated teachers who completed the study (n = 8) differed in their beginning characteristics from the control group participants who completed data collection activities (n = 23).

Results of these investigations are summarized in Table B2. Although differential attrition exists between the treated and control samples over the course of our data collection (treated group attrition rate for the workshop = 47%; control group attrition rate for data collection = 21%), we found that the treated and control teachers remaining in the study (see the columns "Assigned to treatment—completed" and "Assigned to treatment—did not complete" in Table B2) taught in schools with similar environments, had classrooms with similar average student characteristics, and also had similar backgrounds (e.g., similar years of experience). Factors leading to teacher attrition appeared to be similar in both the treated and control conditions. Teachers who failed to complete the study after baseline were more likely to teach a greater proportion of Latinx students and taught in schools with a greater proportion of economically disadvantaged students.

Our understanding of how selection influenced the makeup of our analytic sample guided our decision making about our statistical methods in two ways. First, we used a variety of methods for investigating treatment effects to understand whether the inferences about a treatment effect were sensitive to model selection. Second, the fact that teachers in the treated group who completed the study began with occasional statistically significant higher IQA scores (see Table B2) led to our decision to use a differences-in-differences model to calculate and report treatment effects. The differences-in-differences method ensured that we accounted for any inherent disadvantage for growth in the comparison condition due to starting lower on some IQA measures.

Finally, when exploring estimates of student achievement effects, we examined a variety of models, including using the Heckman method to adjust for factors associated with attrition from the study (i.e., to correct for selection bias or nonrandomly selected samples; Heckman, 1979). This model helped us understand whether treatment estimates were sensitive to factors related to attrition because they account for potential bias that occurred as the result of discontinuing participation. In the first stage of the Heckman model, we predicted whether covariates predicted attrition from the study among both treatment and control teachers (see the Attrition Model section in Table B3).

Four regressors were found to be statistically significant or marginally statistically significant predictors of attrition. In addition to fifth-grade teachers (z = 1.70; p = .088) being more likely to attrit from the study, the other three covariates could have been predicted given the differential attrition across school strata reported in Table B2. Predictors of attrition included the percentage of economically disadvantaged students in the school (z = 3.74; p < .001), the percentage of non-White students in the school (z = 2.23; p = .026), and being a Latinx teacher (z = 1.88; p = .059).

In the second stage of the model, the correction factor (i.e., inverse Mills ratio) is included as a regressor with all of the other statistical controls. As shown in the Reading Achievement Model section in Table B3, findings from the second stage of the Heckman model predict students in classrooms of the treated teachers to score higher by 34.56 points (ES = 0.24; z = 2.24; p = .025).

TABLE B2
Group Differences on Background Characteristics, Including Baseline Teaching on the Instructional Quality Assessment

Covariate	Assigned to treatment—completed	Assigned to control—completed	Treatment vs. control standardized mean difference[a]—completed	Assigned to treatment—did not complete	Assigned to control—did not complete
Student characteristics in student achievement analyses	*N* = 149	*N* = 262	*N = 411*	*N* = 90	*N* = 145
Reading achievement spring 2016	1470.62 (116.07)	1494.51 (130.40)	.18	1434.53	1500.77
Math achievement spring 2016	1484.51[b] (125.26)	1520.42 (143.63)	.25	1465.30	1543.92
Average days absent 2016–2017	8.78 (7.90)	7.54 (6.39)	.17	9.30	7.99
Percentage female	.477	.450	.07	.54	.49
Percentage White	.779	.790	.04	.93	.86
Percentage Black	.134	.088	.28	.05	.06
Percentage Hispanic	.718	.699	.06	.91	.84
Percentage Asian	.013	.031	.54	.01	.01
Percentage Native American	.027	.031	.09	.02	.03
Percentage multiracial	.047	.061	.17	.00	.04
Percentage at risk	.503[b]	.347	.39	.78	.48
Percentage economically disadvantaged	.510[b]	.367	.35	.75	.57
Percentage special education	.107	.088	.13	.11	.12
Percentage handicapped	.087	.107	.14	.07	.15
Percentage bilingual	.054	.042	.16	.18	.07
Percentage migrant status	.007	.008	.08	.01	.01
Percentage gifted	.121[b]	.241	.51	.09	.20
Percentage English learner	.013	.004	.72	.18	.05
Percent limited English proficient	.067	.115	.36	.39	.12

(continued)

TABLE B2
Group Differences on Background Characteristics, Including Baseline Teaching on the Instructional Quality Assessment (*continued*)

Covariate	Assigned to treatment—completed	Assigned to control—completed	Treatment vs. control standardized mean difference[a]—completed	Assigned to treatment—did not complete	Assigned to control—did not complete
School demographics	*N* = 8	*N* = 23	*N* = 31	*N* = 7	*N* = 6
School size	581.12 (128.76)	543.87 (127.36)	.18	589.86	529.00
School percentage White	.18 (.12)	.14 (.07)	.25	.09	.08
School percentage Black	.09 (.06)	.06 (.06)	.17	.05	.03
School percentage Hispanic	.68 (.20)	.75 (.14)	.28	.83	.85
School percentage Asian	.01 (.01)	.01 (.02)	.00	.01	.01
School percentage Native American	.01[c] (.004)	.00 (.003)	.17	.00	.00
School percentage multiracial	.03 (.02)	.02 (.02)	.06	.01	.01
School percentage economically disadvantaged	.64[b] (.19)	.59[d] (.23)	.22	.83	.84
School percentage English learner	.19 (.15)	.26[d] (.22)	.33	.26	.44
School percentage at risk	.54 (.12)	.57[d] (.17)	.18	.62	.70
School percentage mobility	.20 (.07)	.16 (.06)	.51	.18	.19
Teacher survey	*N* = 8	*N* = 20	*N* = 28	*N* = 7	*N* = 6
White	.50	.47	.07	.14	.33
Hispanic	.50	.47	.07	.71	.67
Multiracial	.00	.05	.05	.14	.00
Years of experience	16.15 (10.45)	15.33 (20.60)	.04	13.14	13.67
Bachelor's degree	1	1	.00	1	1
Master's degree	.50	.40[d]	.24	.29	1
Number of English language arts methods courses (categorical 1–6)	3.75 (1.39)	3.05 (1.43)	.47	3.14	4.00
Standard certification	1[c]	.75	.28	.86	.80
Alternate certification	.00[c]	.25	.28	.14	.20

(continued)

TABLE B2
Group Differences on Background Characteristics, Including Baseline Teaching on the Instructional Quality Assessment (*continued*)

Covariate	Assigned to treatment—completed	Assigned to control—completed	Treatment vs. control standardized mean difference[a]—completed	Assigned to treatment—did not complete	Assigned to control—did not complete
Grade 4	.63	.55	.19	.57	.50
Grade 5	.38	.45	.17	.43	.50
Number of students in classroom (2017)	18.50 (2.53)	19.89 (3.74)	.38	19.86	22.60
Content knowledge for teaching English language arts	13.00 (1.93)	12.95 (2.26)	.02	11.86	12.50
Instructional beliefs and practices	*N* = 8	*N* = 20	*N* = 28	*N* = 7	*N* = 6
Teacher-led basic comprehension following Initiate-Respond-Evaluate patterns	3.06 (0.31)	3.08[d] (0.40)	.05	3.23	3.61
Basic-skills comprehension instruction	3.00 (0.82)	3.06 (0.69)	.08	3.21	3.44
Focus on irrelevant factors (independent reading and personal connections)	3.25 (0.57)	3.41 (0.57)	.27	3.14	3.54
Complex texts form the basis for rich text discussion	3.38 (0.70)	3.42 (0.46)	.07	3.45	3.36
Implementation of Questioning the Author	3.53 (0.49)	3.29 (0.35)	.59	3.55	3.53
Students provide meaningful contributions to advance the text discussion	3.32 (0.73)	3.49 (0.49)	.29	3.20	3.54
Instructional Quality Assessment observed at baseline	*N* = 8	*N* = 23	*N* = 31	*N* = 7	
Rigorous text and Questioning the Author implementation composite	2.66 (0.59)	2.53 (0.48)	0.18	2.21	
Teacher facilitation moves composite	2.19[b,c] (0.48)	1.67 (0.43)	1.13	1.48	
Rigor of student contributions composite	2.13 (1.30)	1.61 (0.86)	0.51	1.57	

[a]Where Hedges' *g* was calculated for continuous variables, including all variables with a standard deviation reported, and Cox's index otherwise. [b]Indicates a statistically significant difference in means between the group assigned to treatment and completing the study and the group assigned to treatment and leaving the study. [c]Indicates a statistically significant difference in means between the group assigned to treatment and completing the study and the group assigned to control and completing the study. [d]Indicates a statistically significant difference in means between the group assigned to control and completing the study and the group assigned to control and leaving the study.

TABLE B3
Heckman Model Results

Covariate	Coefficient	Standard error
Attrition model		
Intercept	10.18	7.84
Prior reading achievement	0.00	0.62
Hispanic teacher	−1.18†	0.62
Years of experience	−0.06	0.05
Master's degree	0.37	0.63
Number of English language arts methods courses	−0.06	0.20
Percentage economically disadvantaged students	8.41**	2.24
Percentage non-White students	8.61*	3.86
Grade 4	1.22	1.02
Grade 5	1.83†	1.07
Reading achievement model		
Intercept, γ_{00}	352.65***	96.96
Treated	34.56*	15.40
Stratum 1	−2.58	17.89
Stratum 2	−22.71†	13.98
Grade 4	34.13**	12.91
Grade 5	5.47	9.11
Prior reading achievement	0.61***	0.06
Prior math achievement	0.17**	0.06
Female	15.18*	7.23
Asian	−21.59	24.76
Black	−9.65	23.60
Native American	13.28	20.14
Multiracial	−2.10	19.68
Hispanic	16.86	17.43
At risk	−8.67	15.11
Special education	−7.57	20.71
Handicapped	0.20	20.71
Economically disadvantaged	−13.19	10.19
Gifted	46.46***	13.61
English learner	56.31**	12.44
Limited English proficient	−28.46†	17.13

Note. A cluster option was used to adjust the results for the nested structure of the data. Since Heckman's two-step approach is not compatible with the cluster option, we employed an FIML estimator within a one-step approach. The reported results are very similar to a previous model run with a two-step approach.
*$p < .05$. **$p < .01$. ***$p < .001$. †$p < .1$.

APPENDIX C

Codebook for Elements of Text Discussion Used to Examine the Process of Change

Teacher Initiating Moves[19]

- Each turn could have multiple codes,[20] but do not apply more than one code to the same set of words.
- When assigning multiple codes to a single turn, assign codes in the order in which the words appear in the transcripts.
- Do not consider any responses that follow a teacher turn to code a teacher turn.

Code	Definition of code	Examples	Rules
Partially open-ended question	The question asks students to demonstrate understanding of a particular section of the text. The content of partially open-ended questions includes specific information about, or guidance toward, particular ideas (i.e., provides information or clues as to what is significant about a certain section of the text), OR they are open questions that ask students to grapple with a relatively narrow slice of information within the text OR would invite answers/speculation not grounded in the text (e.g., predictions, talking about personal experiences).	"Why was Salva scared for his friend?" "How do you think this made Salva feel?" Examples of questions considered partially open-ended if they ask about only a narrow slice of information from the text (i.e., are not focused on big ideas or synthesis across a lot of information): • "What can we infer about Nya?" • "What information have we learned about Uncle?"	• Partially open-ended questions do not ask students to synthesize a lot of information. They either include significant clues that guide students to think about the text in specific ways/contain a portion of the answer, OR they only ask students about a narrow slice of information. • If a question is focused on a character's feelings (e.g., "How do you think this made Salva feel?"), code as a partially open-ended question. • *Partially open-ended vs. open-ended gist question:* A partially open-ended question involves the teacher constructing ideas for the students by inserting some significant information. For example, asking questions such as "Why was Salva scared for his friend?" alerts students to the fact that Salva was scared for his friend. A question is also considered partially open-ended if it has too narrow of a focus. For example, if the teacher asks "What information have we learned about Uncle?" after reading a passage that includes information about a lot of other characters and/or ideas, this question would be considered partially open-ended. For open-ended gist questions, the teacher provides students with no explicit information that would guide them to think about the text in any specific way and focuses on big ideas across large chunks of text (e.g., "What is the author telling us here?").
Open-ended gist question	The question asks students to demonstrate understanding of a section of the text as a whole, not simply retrieve bits of information. Cumulatively, the questions help students construct a coherent representation of the text (i.e., big ideas and events).	• "What is the author saying here?" • "What just happened in this section of the text?" • "What did we learn here?" • "What information have we learned about Uncle?" NOTE: This question would be considered an open-ended gist question if it follows a portion of text that contains a lot of important information about Uncle (i.e., information about Uncle represents the big ideas contained in that portion of the text). If information about Uncle only represents a small amount of information among a lot of other important information, it is considered partially open-ended.	• *Open-ended gist vs. partially open-ended question:* Open-ended gist questions ask students to synthesize a lot of information from the text. Open-ended questions that focus on a narrow slice of information (as opposed to big ideas or large swaths of text) or contain a lot of information to guide students' thinking are considered partially open-ended. • *Inference and analysis vs. open-ended gist question:* Gist questions are more about checking students' comprehension of major ideas and events in the text; inference and analysis questions require students to construct knowledge that goes beyond what is directly stated or represented in the text (i.e., require some analysis and higher level inferences). • Questions that ask about a character's feelings (e.g., "How is this making Salva feel?") are considered partially open-ended. • *Literal vs. open-ended gist question:* Literal questions are closed-ended. They are often narrowly worded, such as "What does Nya carry with her?" • *Literal vs. open-ended gist question:* Questions about what a word or phrase means (i.e., vocabulary) are usually literal questions. Occasionally, asking what a phrase means could be a gist question if the importance is placed on what the phrase means in context, not just what it literally means.

Teacher Rejoinder Moves

- When assigning codes, you may need to take into account the previous student's response.

Code	Definition of code	Examples	Rules
Literal uptake	The teacher asks a literal question using a student's response.	"She discovered what he looked like." "Who's the he?" "Responsible." "Responsible. Can anyone think of another word for *responsible*?"	• If the teacher uses a student response but asks a literal question, code as literal uptake. • *Literal question vs. literal uptake:* Literal uptake codes respond to student comments; literal questions are teacher initiated. • *Press vs. literal uptake:* If the teacher asks the same student questions, such as "What page number was that on?" or "Where did you find that?" when the student was literally just reading from the text, and the teacher's purpose is to, for example, get a line or page number to write down, code as literal uptake.
Uptake/ pushback	The teacher uses a student response to extend, deepen, clarify, or elaborate the discussion, OR the teacher challenges the response in order to encourage students to rethink/defend their responses.	"She hadn't killed the dog." "Well, what about hadn't killed the dog?" "Oh, so, what about middle age? So, you're saying that because he was young, he felt that way, but because his parents were older, they felt a different way. Hmm, you think that's true that—do you think age is what determines how he felt about this spring? Courtney? Will that always work?"	• If the teacher uses a student's response (e.g., by paraphrasing or incorporating the response into a question) to extend, clarify, deepen, or elaborate the discussion, code as uptake (usually includes calling on another student). • If the teacher reproduces a student's response verbatim (or close to) and then asks a question related to the response, code as repeat and uptake/pushback (if the teacher calls on another student) or press (if the teacher asks a question to the same student). • If an utterance falls between uptake/pushback and literal, code as literal uptake. • *Collect vs. uptake/pushback:* If more and different answers appear to be the goal, rather than deepening or elaborating a previous student's answer (i.e., uptake), code as collect. • *Press vs. uptake/pushback:* Press is used only when the teacher asks the same student who gave the initial response.

Student Moves

Code	Definition of code	Examples	Rule(s)
Weak link	Students attempt to link contributions to one another but do not show how ideas/positions relate to each other. The student might simply be revoicing or repeating another student's contribution.	"I disagree with Ana" without explaining why or which aspect of Ana's statement the student disagrees with. "I agree with Jay because…" is followed simply by the student's own prepared response, without actually linking to the substance of Jay's response.	• The phrase "I agree/disagree with…" and the connective "because" do not necessarily signal a strong link. Sometimes students use the stem "I agree with ___ because…" in a perfunctory manner. Listen for what comes after the phrase.
Strong link	Students connect their contributions to one other's and show how ideas/ positions shared during the discussion relate to each other. Students elaborate, challenge, or build on one another's ideas.	"I'm not sure what Ana says is right, because I don't see where in the text it says that…." "I agree with Jay because…" is followed by a clear elaboration of Jay's response.	• See above for the rule about distinguishing weak from strong links when students state that they agree or disagree with one another. • Students' building or elaborating on another student's idea must include a unique contribution, not just restate the first idea.
Weak text-based evidence	Students provide inaccurate, incomplete, inappropriate, vague, or trivial evidence from/reference to the text.	"Naya lived a hard life because, in the chapters about her, we learn that she has to do a lot of things for her family."	• Chapter references or general paraphrasings of a large portion of the text instead of specifics are considered weak evidence.

(continued)

Code	Definition of code	Examples	Rule(s)
Strong text-based evidence	Students provide accurate, appropriate, specific evidence from/reference to the text that supports the claim.	"…because it said in the text that Salva did not know the old woman."	• For something to count as evidence at all, it has to be distinct from a student's answer/claim (usually it follows "because" or is marked by "in the text…"). • Specific page or line numbers, specific quotations, or close paraphrasings count as strong text-based evidence.
Weak explanation	Students provide a brief or circular explanation that basically repeats or restates the response or relies on evidence to speak for itself.	"I think Salva was tired because, in the text, it says, 'Salva was tired from the long walk.'"	• Not applicable
Strong explanation	Students provide an elaboration/justification of their answer or of the evidence they selected to support their answer.	"I think Salva was scared for his family because it said he immediately thought of home after the gunfire…"	• Avoid simply coding long responses as strong explanation. Listen for the substance. • Distinguish between the claim and the explanation part of the response. For something to count as explanation, it has to be distinct from students' answer/claim and from the evidence (usually it follows "because…").

APPENDIX D

Tables for Survey Constructs and Findings

TABLE D1
Survey Items Measuring Practices and Beliefs Organized by Construct

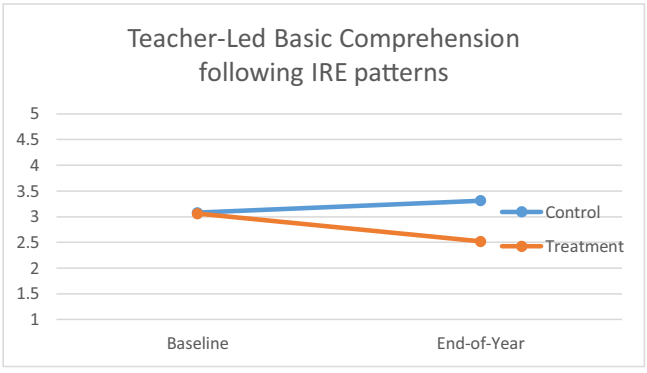
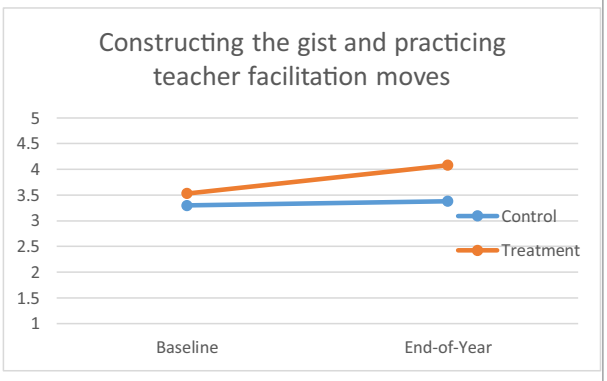
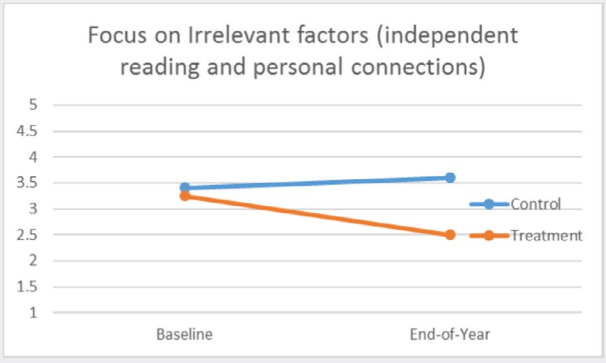
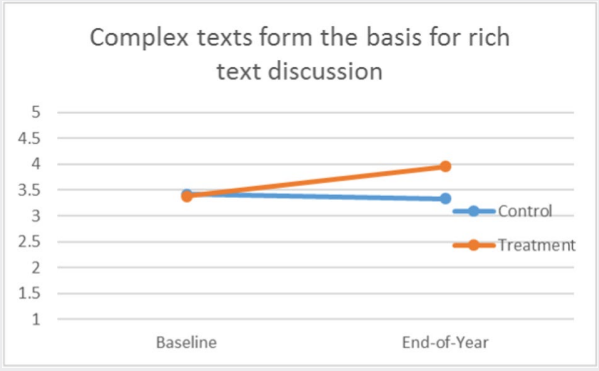
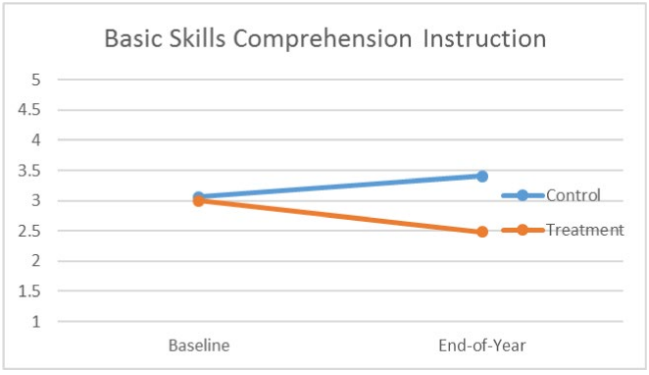
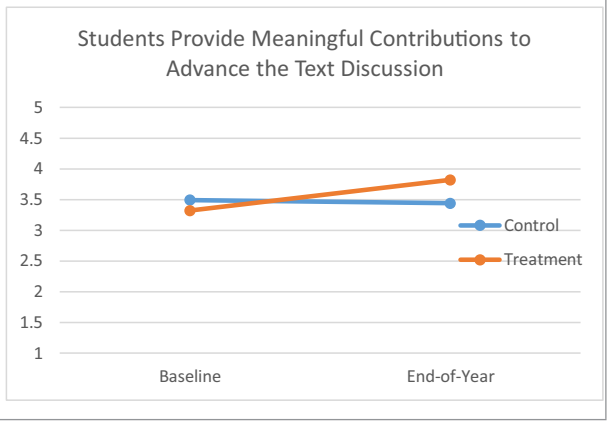
Constructs with hypothesized declines for the treatment group	Beliefs (B) or practice (P)
Teacher-led basic comprehension following Initiate-Respond-Evaluate patterns (a = .60; ES = 1.61; p = *.00)*	
1. It is important that the texts my students and I discuss in class should be consistent with my students' life experiences.	B
2. To improve students' reading comprehension, it is important to ask students to make a prediction about what will happen next in a text.	
3. In our classroom text discussions, my students give short answers (less than a sentence in length).	P
4. In our classroom text discussions, my students only contribute when they know they have the correct answer.	
5. To increase students' reading comprehension, it is important that teachers especially call on students likely to have the correct answer so the whole class will understand the text accurately.	
6. To increase students' reading comprehension, it is important that teachers preteach vocabulary in advance of reading a text.	
7. When I select texts, I choose texts that most of my students can comprehend when reading independently.	
Basic-skills comprehension instruction (a = .80; ES = 1.18; p = *.04)*	
1. It is important that the texts my students and I discuss in class support students to learn a particular comprehension skill or strategy (e.g., identifying cause and effect, predicting).	B
2. To improve students' reading comprehension, it is important to ask very specific questions with correct answers to help students recall important details.	
3. To improve students' reading comprehension, it is important to ask students who, what, where, and when questions to ensure that they comprehended what they read (e.g., "Who are the main characters?").	
4. To improve students' reading comprehension, it is important for teachers to frequently summarize the text for students to make sure that everyone understands the correct meaning.	

(continued)

TABLE D1
Survey Items Measuring Practices and Beliefs Organized by Construct (*continued*)

Constructs with hypothesized declines for the treatment group	Beliefs (B) or practice (P)
5. When I select texts, I choose texts that are similar to the texts found on the state achievement test.	P
6. In our classroom text discussions, my students get confused by questions that invite too many responses (i.e., go beyond the events or facts in a text).	
Focus on irrelevant factors (independent reading and personal connections; a = .66; ES = 1.52; p = *.00)*	
1. It is important that the texts my students and I discuss in class be readily comprehended by most of my students when reading independently.	B
2. To increase students' reading comprehension, it is important that students independently read text.	
3. When I engage students in a text discussion, I avoid addressing factual errors in a student's response so as to maintain that student's self-esteem.	P
4. In our classroom text discussions, my students make predictions about what will happen next.	
Constructs with hypothesized gains for the treatment group	**Beliefs (B) or practice (P)**
Complex texts form the basis for rich text discussion (a = .77; ES = 1.18; p = *.04)*	
1. It is important that the texts my students and I discuss in class be more difficult to comprehend than the texts they read independently.	B
2. It is important that the texts my students and I discuss in class have complex plots (e.g., story contains multiple subplots or is told from multiple characters' perspectives).	
3. It is important that the texts my students and I discuss in class explore problems that do not have clear solutions.	
4. It is important that the texts my students and I discuss in class contain moral dilemmas.	
5. When I select texts, I choose texts that have a complex plot structure (e.g., story contains multiple subplots or is told from multiple characters' perspectives).	P
6. When I select texts, I choose texts that explore problems in the world that do not have clear solutions.	
Implementation of Questioning the Author (a = .61; ES = 1.00; p = *.06)*	
1. To improve students' reading comprehension, it is important to plan questions I will ask in advance rather than just waiting for questions to emerge spontaneously from students during discussions.	B
2. To improve students' reading comprehension, it is important to press an individual student to provide multiple pieces of evidence to support his or her answer.	
3. To improve students' reading comprehension, it is important to ask students to evaluate other students' answers (i.e., if they agree or disagree) and explain why.	
4. When I engage students in a text discussion, I read a text aloud to students.	P
5. When I engage students in a text discussion, I plan questions in advance focused on parts of the text where students might have comprehension problems.	
6. In our classroom text discussions, my students are pressed until they give the correct answer.	
Students provide meaningful contributions to advance the text discussion (a = .87; ES = .89; p = *.10)*	
1. In our classroom text discussions, my students build on one another's assertions by providing new evidence in support of the original claim.	P
2. In our classroom text discussions, my students give more than two-sentence responses.	
3. In our classroom text discussions, my students provide multiple pieces of supporting evidence for a claim.	
4. In our classroom text discussions, my students restate their answers to make sure that they are being understood by the rest of the class.	
5. In our classroom text discussions, my students summarize different ideas voiced by another student to make sure that they understood the student's main point.	
6. In our classroom text discussions, my students connect their answers to another student's ideas.	
7. In our classroom text discussions, my students explain why they agree or disagree with another student's answer.	

TABLE D2
Visual Representations of Differences-in-Differences Models for Survey Constructs

Constructs with hypothesized declines for the treatment group	Constructs with hypothesized gains for the treatment group
	
	
	

APPENDIX E

Growth Analyses for the IQA

To understand growth trajectories for each group of teachers and to better describe overall patterns of growth, we examined the data using hierarchical linear models to identify the best functional form for change (Raudenbush & Bryk, 2002) for teachers on each of the IQA composite scores (i.e., choosing rigorous texts and implementing elements of Questioning the Author, teacher press and facilitation moves, rigor

of student contributions). IQA composite scores for each video were nested in timepoints, and timepoints were nested in teachers in three separate univariate analyses.[21]

These models make use of all available data (112 videos among 31 teachers; n_j = 3.65) to estimate the average growth trajectory and investigate whether growth appears to be different between the treated and control teachers. Our two-level unconditional model for the linear functional form for Questioning the Author implementation can be summarized as follows:

$$\text{QTAimp.}_{ti} = \pi_{0i} + \pi_{1i}\left(\text{Time}_{ti}\right) + e_{ti} \quad (1.1)$$

$$\begin{aligned}\pi_{0i} &= \beta_{00} + \beta_{01}(\text{Online-CFC})_i + r_{0i} \\ \pi_{1i} &= \beta_{10} + \beta_{11}(\text{Online-CFC})_i + r_{1i}\end{aligned} \quad (1.2)$$

where QTAimp._{ti} is the composite score of the three Questioning the Author implementation dimensions (as listed in Table 3); π_{0i} is the baseline Questioning the Author implementation score for teacher *i*; Time_{ti} is the change in time from baseline to the first coached video; π_{1i} is the linear growth slope for teacher *i*; β_{00} is the average baseline Questioning the Author implementation score across teachers; Online-CFC_i is a dichotomous indicator for teacher *i* for participation in Online Content-Focused Coaching (1 = *treated*; 0 = *control*); β_{01} is the difference in baseline Questioning the Author implementation for treated versus control teachers; β_{10} is the average linear growth slope for teachers; β_{11} is the difference in the growth slope for Questioning the Author implementation for treated versus control teachers; e_{ti} is the within-person residual; and r_{0i} and r_{1i} are the between-teacher variance estimates for the intercept at baseline and the linear growth slope, respectively.

APPENDIX F

Piecewise Analyses for Patterns of Growth in Text Discussion Outcomes Across Phases of Online Content-Focused Coaching

Using HLM version 7.03, items were nested in teachers in six separate univariate analyses. Our two-level unconditional model for open-ended gist questions can be summarized as follows:

$$\begin{aligned}\text{ATM}_{open-ended\cdot ti} = {} & \pi_{0i} + \pi_{1i}\left(\text{Workshop Interval}_{ti}\right) \\ & + \pi_{2i}\left(\text{Coaching Interval}_{ti}\right) + e_{ti}\end{aligned} \quad (2.1)$$

$$\begin{aligned}\pi_{0i} &= \beta_{00} + \beta_{01}(\text{Online-CFC})_i + r_{0i} \\ \pi_{1i} &= \beta_{10} + \beta_{11}(\text{Online-CFC})_i + r_{1i} \\ \pi_{2i} &= \beta_{20} + \beta_{21}(\text{Online-CFC})_i + r_{2i}\end{aligned} \quad (2.2)$$

where $\text{ATM}_{\text{open-ended}\cdot ti}$ is the proportion of all teacher moves that were coded as open-ended gist questions at time *t* for teacher *i*; π_{0i} is the baseline proportion of open-ended gist questions at baseline for teacher *i*; Workshop Interval_{ti} is the change in time from baseline to the first coached video; π_{1i} is the linear slope over the workshop interval for teacher *i*; Coaching Interval_{ti} is the change in time from the first coached video to the last coached video for teacher *i*; π_{2i} is the linear slope over the coaching interval for teacher *i*; β_{00} is the average baseline proportion of open-ended gist questions across control teachers[22]; Online-CFC_i is a dichotomous indicator for teacher *i* for participation in Online Content-Focused Coaching (1 = *treated*; 0 = *control*); β_{01} is the difference in the proportion of open-ended gist questions for treated versus control teachers at baseline; β_{10} is the average linear slope for control teachers over the workshop interval; β_{11} is the difference in the change in the average linear slope for the proportion of open-ended gist questions for treated versus control teachers over the workshop interval; β_{20} is the average linear growth slope over the coaching interval; β_{21} is the difference in the change in the average linear slope for the proportion of open-ended gist questions for treated versus control teachers over the coaching interval; e_{ti} is the within-person residual; and r_{0i}, r_{1i}, and r_{2i} are the between-teacher variance estimates for the intercept at baseline, the workshop interval, and the coaching interval, respectively.

APPENDIX G

Excerpts From Text Discussions at Baseline, After the Workshop, and at the End of the Year

Illustrative Example of Evolving Text Discussions in One Treated Classroom

Transcript excerpt from:	ATM code
Baseline (September)	
Teacher: So, what is the main problem in the story? Who would like to share? Ben, what do you think?	Partially open-ended question
Ben: I think the problem is when Charlie gets the nightmares, and he tries to get them away from him.	Weak explanation
Teacher: OK, I like that one. Anybody else think something different? Nina?	(Evaluate)/collect
Nina: I think the problem is his nightmares are getting worse and worse and…	NC
Teacher: I wonder why.	Redirect
Elena: Because Charlotte and the witch…	NC
Teacher: OK. Are they the same person or different people, Charlotte and the witch?	Literal uptake
Elena: I think a little bit the same people.	NC
Teacher: You think they're the same person, OK. Javier?	Repeat/collect
Javier: I think that it's about his mother and Charlotte because he doesn't want to leave his mother behind, and his father is saying leave her behind and—	Weak text-based evidence
Teacher: And what happened to his mother?	Literal uptake
Javier: His mother died. That is also one problem.	NC
Postworkshop (January)	
Teacher: [Reads portion of story aloud]	NC
Teacher: What happened now? Javier?	Open-ended gist question
Javier: While Scho was trying to distract them, they want to take a little break so Scho won't say that again. He will think it, but now it's going to go annoying, and they're ticking them off.	Strong explanation
Teacher: Ah, I like how you said "and they're ticking them off." OK. So, is the distraction working? I don't know. Can you add on to that, Crystal, what Javier told us?	(Evaluate)/invitation to link
Crystal: Well, they—he is ticking them off. But at the same time, he's also really annoying them by singing and keep on saying that he told them to say that.	Weak link
Teacher: So, how are Monk and Glennie taking all of this? Destiny?	Partially open-ended question
Destiny: They're being annoyed, and because Scho wanted them to stop the game, so they did. So, he can get more attention because he's not playing.	Strong explanation
Teacher: Ah, I like what she said about getting more attention.	(Evaluate)/repeat

(continued)

Illustrative Example of Evolving Text Discussions in One Treated Classroom (*continued*)

Transcript excerpt from:	ATM code
Postcoaching (end of the year)	
Teacher: How do you feel about what Mario said? Do you agree or disagree? Aaron?	Invitation to link
Aaron: I agree with him because last time Novello tried kissing her, and Delly punched him.	Strong link
Teacher: OK. Damian?	Collect
Damian: I agree with Javier and kind of Aaron because he did try to kiss her, and when he tried, that's when he like—he's getting more close to her because now she's starting to get mad, and he's thinking to him he's probably teasing him. So, that's probably why he keeps teasing her, because he wants her to tease him.	Strong link
Teacher: OK. Cathy?	Collect
Cathy: I disagree with Daniel.	NC*
Teacher: Why?	Press
Cathy: I don't think he's trying to get Delly to tease him. I just think that—	NC
Teacher: Keep going. You don't think he's trying to get Delly to tease him. What do you think he's trying to do?	Press
Cathy: I guess get her to like him somehow.	Strong link
Teacher: Why do you think that?	Press
Cathy: Because in a couple chapters ago, almost the beginning of the book, it said he loved Delly.	Strong evidence
Teacher: OK.	NC
Ben: I agree with Cathy because in some way he's bugging her because he likes her, so he's just trying to bug her.	Strong link
Teacher: OK.	NC
Maya: I disagree with Cathy because he said he missed her—what was it again?	NC*
Sofia: Punches?	NC
Teacher: Torturous touch.	Provides information
Maya: Yeah, that. I'm sure he's trying to get her to hit him or tease him or be mean to him, but she hasn't been doing it lately. She's been counting, so she's…	Strong explanation
Teacher: OK.	NC
Robert: I think he just wants her attention.	Strong link
Teacher: Why do you think that?	Press
Robert: Because who does that? When someone's teased, or from my personal experience with [my sister], whenever she teases, it's supposed to get my attention and annoy me. I think that's why [Novello's] doing that, so [Delly] says something to him. I think he just wants her attention.	Strong explanation

Note. ATM = analyzing teaching moves; NC = not codable; NC* = the student response would fit the definition of a weak link if left by itself, but because the student continues the response in immediate subsequent turns (after pressed by the teacher to elaborate), only the contribution across all turns is coded. (Evaluate) is not a code in the ATM codebook but was added here to help describe the change in text discussions over the course of the year away from an Initiate-Respond-Evaluate pattern and toward facilitation of students' reasoning about the text. The text discussed at baseline is *Nightmares!* by Jason Segel and Kirsten Miller (2014), the text discussed postworkshop is *A Game of Catch* by Richard Wilbur (1994), and the text discussed postcoaching is *True (…Sort Of)* by Katherine Hannigan (2011).