Educational Evaluation and Policy Analysis
September 2020, Vol. 42, No. 3, pp. 439–466
DOI: 10.3102/0162373720940699
Article reuse guidelines: sagepub.com/journals-permissions

Mathematics Coaching for Conceptual Understanding: Promising Evidence Regarding the Tennessee Math Coaching Model

Jennifer Lin Russell
Richard Correnti
Mary Kay Stein
University of Pittsburgh
Ally Thomas
UPMC Health Plan
Victoria Bill
Laurie Speranzo
Institute for Learning

Rigorous college-and-career readiness standards require significant shifts in typical mathematics instruction. Many schools and districts employ coaches to support instructional changes. Although there is evidence that coaching programs can support teaching improvement, research has yet to identify high-leverage coaching practices. In collaboration with a network of state leaders and coaches, our research team refined a model for math coaching and documented the practices coaches employed in one-on-one work with teachers. Analysis of videotaped coaching conversations and teaching events suggests that model-trained coaches improved their capacity to use a high-leverage coaching practice—deep and specific prelesson planning conversations—and that growth in this practice predicted teaching improvement, specifically increased opportunities for students to engage in conceptual thinking.

Keywords: *mathematics education, educational policy, professional development, longitudinal studies, hierarchical linear modeling*

Policymakers and educational leaders around the country are grappling with how to support mathematics teachers in shifting the focus of their teaching from the demonstration of algorithms for solving routine problems to the development of students' conceptual understanding. This shift has been fueled by a state-led movement to prepare all of America's students to graduate from high school college- and career-ready, a movement associated with the National Governors Association's release of the Common Core State Standards—Mathematics (CCSM-M). When coupled with aligned assessments, these standards place demands on teachers to use more cognitively challenging instructional tasks and to orchestrate productive classroom discussions that surface and build on students' thinking and reasoning, all in service of building students' conceptual understanding of mathematical ideas. This kind of teaching is a significant change in practice for many elementary- and middle-grade teachers (Hiebert, 2003; Stein & Meikle, 2017).

In response to the ambitious goals for instructional change promoted by the new standards for mathematical practice, states and districts have sought to refine or elaborate their instructional guidance infrastructures. Although traditional policy approaches to teaching improvement have focused on the introduction of new curricula and aligned professional development, there is a growing consensus in the educational field that these strategies alone are insufficient to produce the instructional change necessary to ensure that all students can meet ambitious standards. Consequently, many schools and districts have created instructional coaching roles to complement more traditional professional development efforts as the focus on college and career readiness has taken hold in the field (Desimone & Pak, 2016; Domina et al., 2015). Although roles vary in implementation, typically instructional coaches are charged with creating intensive, job-embedded learning opportunities for teachers through strategies such as offering ongoing workshops, leading professional learning communities, and working one-on-one with teachers to "coach" their preparation for and execution of lessons.

The creation of instructional coaching positions has the potential to be a high-leverage district strategy for supporting substantive changes in teaching practice. Prior research suggests that, when well designed, coaching initiatives can produce measurable gains in teaching improvement and student achievement (Allen et al., 2011; Biancarosa et al., 2010; Blazar & Kraft, 2015; Bryk et al., 2015; Campbell & Malkus, 2011; Foster & Noyce, 2004; Kraft et al., 2018; Matsumura et al.,2010, 2012, 2013; Neuman & Cunningham, 2009; Powell et al., 2010; Sailors & Price, 2010). Specific to mathematics instruction, Campbell and Malkus's (2011) randomized-controlled trial found that coaches positively affected elementary student mathematics achievement (Grades 3–5), particularly after coaches gained experience and skill through extensive professional development.

However, the investment in coaching nationwide has likely not yielded its full potential (Russell et al., 2017). Research suggests that coaching programs have variable outcomes due in part to implementation challenges such as insufficient coach training, guidance, and support (Gallucci et al., 2010; Kraft et al., 2018; Matsumura et al., 2009). For example, coaches are routinely selected for their teaching excellence but may know little about the practice of coaching. Despite the promising results from some coaching studies, far less is known about the specific features of coaching programs, including high-leverage coaching practices that can support teaching for mathematical understanding. Coaches take on a variety of roles that, theoretically, could support improvements in teachers' instruction: providing regular professional development sessions, leading grade-level meetings, one-on-one coaching, developing teachers' content knowledge, and so on. Moreover, *within* each of these roles, various practices could be embedded. Few studies have sought to explore how specific coaching practices are associated with changes in teachers' instruction.

To address this gap in the field's understanding of coaching as an intervention for driving teaching improvement, our collaborative project team, including researchers, professionals who design and implement professional development for educators, a state education agency, and a network of mathematics coaches, iteratively refined a model for mathematics instructional coaching. The model aimed to be a resource for state and local leaders as they develop coaching programs to support instructional improvement at scale. Utilizing a design-based implementation research (DBIR) approach, organized by improvement cycles, we carefully documented the practices coaches employed in their one-on-one work with teachers and identified practices associated with improved teaching. In subsequent improvement cycles, coaches received training in the high-leverage coaching practices that emerged from our improvement research.

In this article, we identify a high-leverage coaching practice—deep and specific conversations about mathematics content, pedagogy, and student thinking occurring in prelesson planning conferences—and explore whether coaches improved the uptake of this practice over time. In addition, we explore the extent to which teachers who participated in these prelesson planning conferences improved their capacity for supporting students' mathematical reasoning and problem-solving, and whether improvements in coaching predicted growth in teaching. In so doing we contribute to the field's understanding of the relationship between coaching practice and teaching improvement, which can in turn provide guidance to states and districts developing instructional coaching initiatives.

Literature Review and Conceptual Framework

Teaching for Mathematical Understanding

College- and career-readiness focused standards require that teachers teach for mathematical understanding. Prior research has demonstrated that classrooms in which students productively struggle to complete tasks that are just beyond their reach, and in which explicit attention is paid to underlying concepts and ideas, are associated with students' development of conceptual understanding (Hiebert & Grouws, 2007). Support for this association comes from mathematics education research (Boaler & Staples, 2008; Hiebert & Grouws, 2007; Otten & Soria, 2014; Stein & Lane, 1996; Stigler & Hiebert, 2004) and from cognitive science (e.g., DeCaro & Rittle-Johnson, 2012; Kapur, 2014; Schwartz & Martin, 2004).

Mathematics education research also suggests a set of practices that teachers who effectively use a student-centered, discussion-based approach engage in when they plan for and enact high-quality lessons (Stein et al., 2008). These include anticipating possible student thinking pathways (Smith et al., 2008), designing finely tuned goals regarding what students will know and understand (Hiebert et al., 2018), setting up challenging tasks (Jackson et al., 2013), encouraging students to use multiple perspectives/strategies and explain their responses or solution strategies (Tarr et al., 2008), and promoting student engagement with each other's ideas during whole class discussions (Franke et al., 2015; Webb et al., 2014). Similar practices were found to be conducive to effective student-centered instruction in cognitive science research, for example, using complex mathematical problems and asking students to generate multiple solutions (Kapur, 2012, 2014) and connecting student thinking to canonical solutions (Schwartz & Martin, 2004).

All of this suggests that we know the features of instruction that—if well executed in the classroom—should lead to improved development of conceptual understanding among students. As such, these features become practices on which coaches might focus as they support the improvement of mathematics teaching.

Coaching Teachers to Teach for Conceptual Understanding

We situate our research and design work within one-on-one coaching that assists teachers as they plan a lesson, teach the lesson, and then reflect on it during a postlesson debrief. As just noted, we have relatively strong evidence pointing to specific features of instruction that enhance students' *conceptual understanding* of mathematics. Getting teachers to learn these practices and supporting them as they refine their skills represents a significant challenge for schools, districts, and ultimately state departments of education. Workshop-based, short-term trainings are the most common form of teacher professional development in the United States. (Darling-Hammond et al., 2009; Gravani, 2007; Webster-Wright, 2009). When well designed, these kinds of training sessions have value for increasing teacher knowledge, which is important for improving teaching practice. However, changing instructional practice is more complex than simply increasing teacher knowledge, and thus short term and episodic workshops are likely insufficient to produce substantive changes in teaching (Kennedy, 2016; Opfer & Pedder, 2011). In fact, research on professional development programs suggests it has highly variable results, with most programs showing limited effects on teaching practice (Hill, 2009; Kennedy, 2016). In response to these findings, researchers and reformers have identified suspected principles of more effective professional development opportunities for teachers, calling for learning opportunities to be intensive, ongoing, job-embedded, active, and content specific (Desimone, 2009; Desimone & Garet, 2015; Garet et al., 2001). Instructional coaching programs may be a way to embed such learning opportunities for teachers in schools and systems.

Research supports this assertion, producing evidence that instructional coaching can contribute to substantive change in teaching and improvement in student learning outcomes (cf., a meta-analysis on coaching by Kraft et al. [2018] and Campbell and Malkus's [2011] experimental study). Some quasi-experimental and observational studies provide further promising evidence that coaching programs, when well designed, can be a promising intervention worthy of scale (Foster & Noyce, 2004; Garet et al., 2011; Killion, 2012; Mangin & Dunsmore, 2014; Neufeld & Roper, 2003; Polly, 2012). However, the research on coaching also shows that effects on teaching and student learning are variable. Blazar and Kraft's (2015) experimental evaluation of a coaching program found that coaches varied significantly in their effectiveness at improving teachers' instructional practice; while the overall effect of coaching was nearly a full standard deviation in the study's measure of teaching practice, individual coaches ranged from significant positive effects to negative effects.

Two implementation challenges have emerged as potential explanations for these variable results. First, coaches often get insufficient training, guidance, and support to enact their roles in ways that contribute to teacher professional learning and instructional change (Gallucci et al., 2010; Kraft et al., 2018; Matsumura et al., 2009). Such support is important; although coaches are often selected because they are good teachers of mathematics to students, they may know little about how to support teachers as learners. Second, what constitutes "coaching" varies significantly, as does coaching effectiveness. For example, the amount of time teachers work with coaches varies considerably across coaching intervention studies, though the effects of coaching dosage are not always significant (Kraft et al., 2018). Coaching effectiveness also varies, and may be, in part, associated with the number of teachers that coaches work with at one time (Atteberry & Bryk, 2011). The focus of coaching varies, ranging from content-specific teaching practices to general instructional practice to behavior management (Blazar & Kraft, 2015). All of this suggests there is a need for research-based coaching models that more explicitly grapple with the challenges of coaching program implementation (Kraft et al., 2018).

Conceptual Grounding for the Tennessee (TN) Math Coaching Approach

Although there are a number of coaching models being promoted in education, the core features are not often well explicated or are ambiguous (Gallucci et al., 2010). For example, Poglinco and colleagues (2003) describe the coaching model in the America's Choice comprehensive school reform model as a form of technical coaching used to transfer new teaching practices into teachers' regular repertoires. However, findings from their study of the enactment of the coach role suggest role expectations were ambiguous, leading to variation in the practices employed by coaches.

To better understand how coaching contributes to teacher development, it is critical that the field articulates and identifies coaching practices. There is some small scale, qualitative research that has investigated coaching practices. For example, Gibbons and Cobb's (2016) investigation of one coach's practice identified five aspects of coach's planning practices when working one-on-one with teachers trying to promote ambitious instructional practices. Other studies have explicated important insights about how coaches build relationships and rapport with teachers that enable learning (Killion, 2008). However, these qualitative studies have not traditionally linked observations about coaching practices with analysis of teaching.

Given the field does not know much about what coaching practices contribute to teaching improvement, our research and development activities were grounded in broader findings about teaching development and the mechanisms whereby coaching might support improved teaching practice. Our model is distinctive in its focus on one-on-one coaching that targets planning, enacting, and reflecting on a specific lesson, as well as its focus on core disciplinary teaching practices. In other words, it specifically focuses on building teacher capacity to enact rigorous mathematics tasks that provide opportunities for student reasoning about mathematics concepts.

Our initial specification of key coaching practices drew on prior research that emphasizes the importance of planning for rigorous instruction (Lewis, 2002; Lewis & Tsuchida, 1997; Stein et al., 2008). Teaching for conceptual understanding requires that teachers select, adapt, or create instructional tasks that will challenge students to think, reason, problem solve and apply previously learned skills to novel situations.

These kinds of tasks are challenging for teachers to learn to enact well because they follow a much-less predictable route than do conventional lessons (Stein et al., 1996). Teachers must be prepared to deal with the wide range of student strategies and responses that typically occur, to make sense of them, and gently coax them toward the goals of the lesson (e.g., Lampert, 2001).

Lesson planning in which teachers set goals, select tasks aligned with those goals, anticipate student responses and contributions, and identify how to make productive use of them makes such teaching more manageable, focused, and productive (Smith et al., 2008; Stein et al., 2008). It does so by helping teachers to plan in advance for the improvisational aspects of responding to students while at the same time guiding the class toward the goals of the lesson (cf. Fennema & Franke, 1992; Gravemeijer, 2004).

Prior research suggests teachers' uptake of this kind of planning may be facilitated in collaborative settings such as professional learning communities and one-on-one coaching. For example, in the lesson study routine, teachers engage in substantive collaborative planning which includes considering long-term goals for student learning, studying existing instructional materials, planning a "research lesson" (including anticipated student contributions), teaching and observing the lesson, and collaboratively analyzing data from it (Fernandez & Yoshida, 2004; Lewis, 2002; Lewis & Tsuchida, 1997, 1998). Beyond the context of lesson planning exclusively, Coburn and colleagues (Coburn et al., 2012; Coburn & Russell, 2008) identified teachers engaged in high "depth" conversations as those that took up substantive issues related to teaching and learning. These examples provided inspiration as we conceptualized the way teachers' interactions with coaches could contribute to their capacity to teach for conceptual understanding.

Drawing on mathematics instructional research, we conceptualize deep and specific coaching conversations as those that focus on the interactions between teachers, students, and mathematics, (not solely, for example, on what the teacher will do). Using the instructional triangle (Cohen et al., 2003), coaching is framed by the view that opportunities for student learning are constituted not by any one of these components alone, but rather by their interaction. As such, we posit that coaching sessions should support teachers in attending to the interaction of pedagogy, student thinking, and the mathematics as they plan lessons.

All of the above has led us to hypothesize several suspected principles of effective coaching conversations about the teaching of mathematics: Such conversations are *deep* in substance (as opposed to focusing on superficial features), *specific* in regards to what they target, and, finally, they occur in the context of the *instructional triangle*. Adherence to these principles during prelesson planning conferences constitutes a key coaching practice in our model: deep and specific conversations about mathematics, pedagogy, and student learning (the instructional triangle).

Present Study

The purpose of the present study is to examine whether there is evidence that coaches and teachers who participated in the TN Math Coaching Project improved their coaching and teaching in ways that align with the TN + Institute for Learning (IFL) Math Coaching Model's core principles. Understanding how the specific features of coaching specified in the TN + IFL Math Coaching Model contribute to teaching improvement provides the kind of concrete, empirically grounded guidance for instructional leaders in school and districts who are designing and supporting instructional coaching initiatives. Specifically, we explored whether coaches improved in their capacity to have deep and specific prelesson planning conversations with teachers and whether teachers improved their capacity to provide opportunities for students to reason about mathematics. While the data and analyses are primarily descriptive, we explored the extent to which growth in coaching predicted growth in teaching, as an indicator of the model's promise of efficacy. The following research questions guided the study:

Research Question 1 (RQ1): To what extent did coaches trained in the TN + IFL Math Coaching Model improve the depth and specificity of prelesson planning conversations with teachers over time? **(1a)** To what extent is there variation in the growth of the depth and specificity of prelesson planning conversations across coach–teacher pairs over time?

Research Question 2 (RQ2): To what extent did teachers coached by coaches trained in the TN + IFL Math Coaching Model improve students' opportunities to engage in conceptual thinking over time? **(2a)** To what extent is there variation in students' opportunities to engage in conceptual thinking over time?

Research Question 3 (RQ3): Are the depth and specificity of the teachers' prelesson coaching conversations with their coach related to growth in their teaching?

We hypothesized that improvements in the depth and specificity of the prelesson planning conversations that occurred between teachers and coaches prior to teaching events would contribute to improvement in the enactment of lessons. Specifically, by guiding teachers in deep and specific conversations about what mathematics their students should learn in the upcoming lesson and how teachers might scaffold students' learning of that mathematics, coaches will equip teachers with the knowledge and skills to not only select rigorous, high-level tasks, but also to maintain the high-level of thinking, reasoning, and problem-solving that high-level tasks are designed to elicit and support.

Project Context

The TN Math Coaching Project, funded by the Institute of Education Sciences, is a partnership between the Tennessee Department of Education (TDOE) and the University of Pittsburgh aimed at improving the in-service training of Grades 3–8 mathematics teachers as a route to improving the math achievement of all Tennessee students. University of Pittsburgh researchers from the Learning Research & Development Center (LRDC) partnered with scholar practitioners from the IFL, an outreach division of LRDC that provides professional development grounded in research-based practices. Our work had two primary goals. First, we tested and iteratively refined a model for mathematics instructional coaching that is designed to be a resource for schools and districts throughout the state as they support the transition to teaching that is aligned with rigorous, college-and-career ready mathematics standards. Second, we sought to catalyze a network of highly trained coaches that can serve as instructional leaders throughout the state.

Our work to test and refine a model for mathematics instructional coaching employed a DBIR approach (Penuel et al., 2011). DBIR is an appropriate method for organizing research–practice partnerships aimed at addressing complex problems of practice. It is rooted in problems of practice experienced by practitioners and policy actors and employs a research-based approach to design and test interventions using systematic data collection and analysis strategies. Our collaborative work initially started with the TDOE identifying a problem of policy practice: advancing the state's mathematics instructional improvement goals related to ensuring all teachers were teaching for conceptual understanding, drawing on the policy levers available to a state agency. We drew on prior research and practice-based knowledge that our IFL colleagues acquired through their professional development work with coaches to specify key coaching practices and a method for training coaches to enact these practices.

Through five iterative cycles, we trained coaches to enact specific practices, documented the enactment of the practices with two "partner teachers" per coach, analyzed data to understand the enactment of practices, and then refined training and guidance to coaches for the next cycle. At the end of each school year, after the third and fifth coaching cycles, we also explored trends in coaching and teaching practice and how they were related. In the following sections, we provide more detail about our sample, data collection procedures, and analytic approach.

Participants

Our sample of 32 coaches was selected through a competitive process. Our partners from the state department of education distributed an announcement about our coaching project to all school districts and in a number of educator forums throughout the state. Interested coaches submitted a written application, which included a statement of interest in the project, a statement of experience and effectiveness as a coach, and a performance task in which applicants identified a high level mathematical task and learning goals for a lesson with a given focus and anticipated how students might solve the task and ways a teacher could support students' conceptual understanding. We received 62 complete applications. Applicants then participated in a performance-based oral interview conducted by our IFL colleagues, which included analysis of two written scenarios of mathematics instruction and role-playing a coaching interaction related to each scenario.

We utilized a rubric to score coach performance on the written and oral portions of the application process. The rubric had five dimensions: belief that all students and teachers can learn and improve; evidence of content knowledge; attention to student thinking and reasoning; coach disposition as a learner; and communication effectiveness. In the end, we selected 32 coaches that represented variation in prior experience and training, coaching context (e.g., urban, suburban, or rural), and the construction of formal coaching roles (e.g., school-based vs. district-based). All coaches were in full-time instructional coaching positions; in other words, they did not have teaching responsibilities. Coaches were paid an annual stipend of US$3,000 to compensate them for time spent in extra-duty responsibilities related to gathering and transferring data on coaching and teaching to the research team. Specifically, coaches videotaped their coaching conferences and teacher lessons associated with coaching cycles and managed the transfer process. In addition, they received a tablet computer for use in videotaping conferences and lessons that they kept at the conclusion of the study and were reimbursed for travel costs associated with attending network meetings. At the end of the first year of the project, two coaches decided not to continue into Year 2. One coach retired and the other found that disorder in her coaching context did not enable her to devote the time necessary to one-on-one coaching. We contacted two additional coaches from the original applicant pool and invited them to join in Year 2.

Our original 32 coaches were asked to identify two "partner teachers" that would engage in intensive coaching cycles and participate in data collection for the study. Coaches were instructed to select partner teachers that were willing to participate, had room for teaching improvement, and taught students in Grades 3 through 8. After analyzing teaching practice across the first year, we found that 31% of partner teachers entered the project with relatively strong teaching practices and maintained a high level of teaching quality throughout the year. To learn more about teaching improvement, we asked each coach to invite one of their two partner teachers to remain in the study that had the greater need for teaching development and to invite one new teacher to join the project who had need for teaching improvement. In some cases, coaches were not able to replace a teacher or had to replace both of their Year 1 partner teachers due to specific circumstances (e.g., no other available teachers in the target grades in small elementary schools). In total, our sample includes 103 partner teachers: 41 partner teachers who participated in Year 1 only, 38 who participated in Year 2 only, and 24 who participated in both Years 1 and 2. Partner teachers were paid for engagement in data collection activities, such as US$20 to US$50 (depending on time required) for the completion of surveys on experiences with coaching and their teaching practice.

Data Collection and Measures

Coaches were trained in the TN + IFL Math Coaching Model during three 2-day face-to-face sessions per year with monthly webinars for discussion and reflection in between. The vast majority of coaches attended all six face-to-face workshops, totaling approximately 48 hours of training spread across 2 years. Twelve 1-hour webinars were offered, but attendance was less consistent in these session. Consequently, most coaches received approximately 55 hours of training across two school years.

The face-to-face workshops and webinars were conducted as full group sessions, with opportunities for small group engagement, and included activities aimed at building coaches understanding of high-leverage coaching practices such as deep

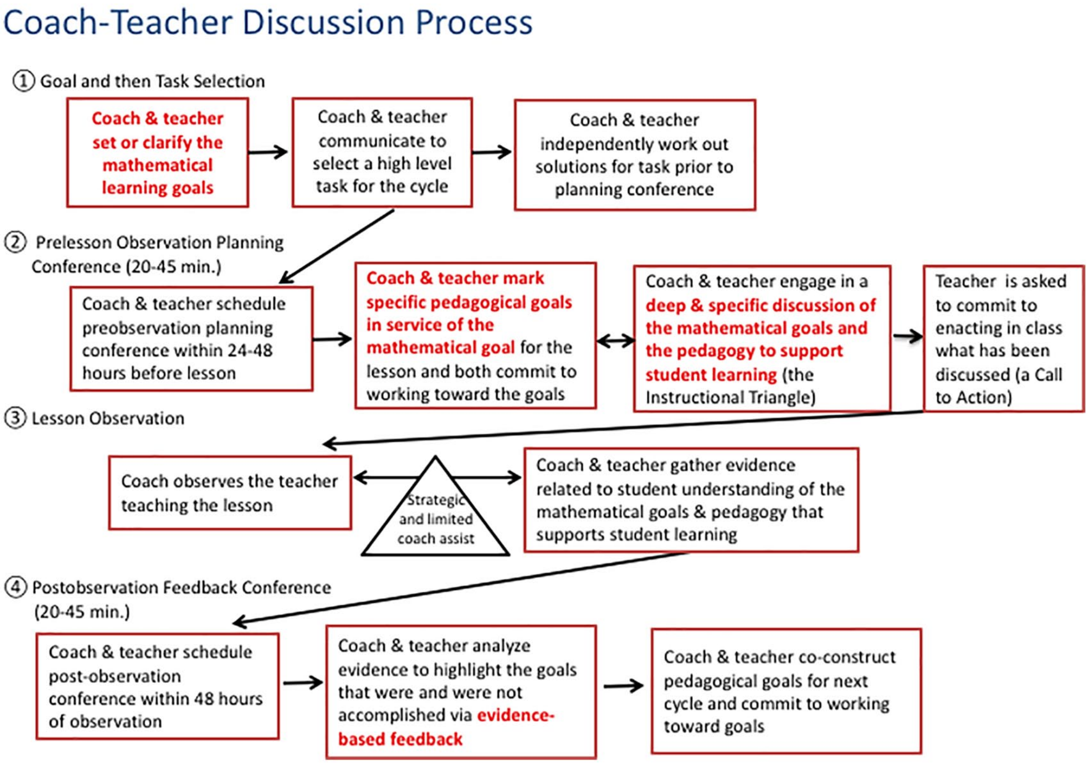
Coach-Teacher Discussion Process
① Goal and then Task Selection
Coach & teacher set or clarify the mathematical learning goals
Coach & teacher communicate to select a high level task for the cycle
Coach & teacher independently work out solutions for task prior to planning conference
② Prelesson Observation Planning Conference (20-45 min.)
Coach & teacher schedule preobservation planning conference within 24-48 hours before lesson
Coach & teacher mark specific pedagogical goals in service of the mathematical goal for the lesson and both commit to working toward the goals
Coach & teacher engage in a deep & specific discussion of the mathematical goals and the pedagogy to support student learning (the Instructional Triangle)
Teacher is asked to commit to enacting in class what has been discussed (a Call to Action)
③ Lesson Observation
Coach observes the teacher teaching the lesson
Strategic and limited coach assist
Coach & teacher gather evidence related to student understanding of the mathematical goals & pedagogy that supports student learning
④ Postobservation Feedback Conference (20-45 min.)
Coach & teacher schedule post-observation conference within 48 hours of observation
Coach & teacher analyze evidence to highlight the goals that were and were not accomplished via evidence-based feedback
Coach & teacher co-construct pedagogical goals for next cycle and commit to working toward goals

FIGURE 1. *Coach–teacher discussion process.*

and specific conversations about the instructional triangle and providing evidence-based feedback. For example, regular activities in the face-to-face workshops included direct instruction in specific coaching practices, coaching simulations/role-playing, collaborative analysis of videotaped coaching sessions, and/or transcripts of coaching interactions, and presentations from the research team sharing the results of analyses of the data coaches shared (videos of coaching sessions and teachers' instruction). During the data-based sessions, coaches and the research team engaged in conversation to make sense of and interpret trends in the data. Most of the training was provided in whole group settings; however, at three strategic points in time, the project team provided individualized written feedback to coaches about their uptake of key coaching practices by annotating a transcript of one of their coaching conversations. Finally, some of the coaches reached out to the professional development providers with informal questions and requests for advice.

Between meetings, coaches were asked to apply what they learned by conducting formal coaching cycles with two partner teachers using the Coach–Teacher Discussion Process. Figure 1 describes the steps in the process. The statements depicted in boldfaced text in Figure 1 highlight three key coaching practices, which have been identified through our iterative work examining variation in coaching interactions through the lens of theory about teaching development. Coaches completed the Discussion Process with each of their partner teachers 3 times in Year 1 of the study (2014–2015) and twice in Year 2 of the study (2015–2016), each corresponding with a project improvement cycle. For each cycle, coaches worked with teachers to plan and enact a lesson that included a high cognitive demand task of their choosing. Given that teachers were in schools and districts that used a variety of curricula, in most cases, the tasks were selected by coaches and teachers from a repository provided by the TDOE and created by the IFL. To capture intensive data on coaching practice, each enactment of the Discussion Process was documented by coaches who gathered: videotapes of prelesson planning conferences, lessons, and postobservation feedback conferences; teacher and coach planning and reflection notes; and artifacts such as the instructional tasks. Coaches shared these data sources with the research team by uploading them to a shared folder.

In addition, we sought to collect representations of each partner teachers' instruction that was not directly associated with a coaching cycle. At the beginning of Year 2 of the study, prior to beginning to work with their coach, we asked partner teachers to videotape a lesson that represented their typical mathematics instruction (Year 2 baseline lesson). At the end of the school year, we asked partner teachers to videotape and share a lesson representing their instruction, that they did not work with a coach to plan or implement (Year 2 postcoaching lesson).

To answer our research questions about changes in coaching and teaching, we identified measures that would track expected changes that related to the training provided to coaches and the instructional philosophy promoted through our coaching model. An overview of our measures is provided in Table 1, and a more complete description in the following sections.

Deep and Specific Prelesson Planning Conversations. Our analyses of coaching began with an exploratory approach. Framed by the guidance we provided to coaches in each cycle, we engaged in a mix of inductive and deductive coding of videos and transcripts of coaching conversations, to surface seemingly productive features of coaching interactions documented with video. Our deductive exploration drew on the list of coaching practices culled from our IFL partners' practice-based insights. Over time, our team came to consensus around three key coaching practices depicted in red text in Figure 1: deep and specific conversations about the instructional triangle, goal setting, and evidence-based feedback. At the end of the first year of the study, we explored the relationship between enactment of each coaching practice and trends in partner teachers' instruction. Given resource constraints, and emerging evidence of its importance, we invested our analytic capacity in developing a measure of deep and specific prelesson planning conversations and pursued it as a specific focus of inquiry. As this coaching practice occurred in the context of prelesson planning conferences (conducted as part of the planning, enacting, and reflecting coaching routine), we refer to it as "preconference depth and specificity" for short.

Based on exploratory analyses, we developed and iteratively refined a rubric for scoring the enactment of key components of deep and specific prelesson planning conversations. The four key components of this construct were: (a) the appropriateness of mathematical content in the task for the grade level, (b) the depth at which multiple solution paths for the task were discussed, (c) whether specific questions to advance the conceptual goals of the lesson were identified and discussed at depth, and (d) the depth of discussions about the specific math content and goals of the task (Figure 2). Six independent coders applied the rubric to all prelesson planning conferences documented in the five coaching cycles. In addition, we calculated a composite Preconference Depth and Specificity score for each preconference by taking the mean of the four items.[1] Given this measure was developed in the context of the project, the coding proceeded in an iterative fashion. Initially, coders scored the same transcripts using the rubric and discussed impressions to build consensus. Once they were reliably scoring transcripts consistently, the coders continued to meet weekly, scoring one common transcript in addition to 3 to 5 others. Again, the weekly meetings resolved any discrepancies in the common transcript and clarified expectations. Approximately 5% of transcripts were scored via consensus.

Students' Opportunities to Engage in Conceptual Thinking. To identify changes in teaching practices throughout the years, we coded the classroom videos from each coaching cycle and

TABLE 1

Coaching and Teaching Measures

Measure	Definition	Metric
Preconference depth and specificity composite (average of the following four measures):		
Appropriate math content	The extent to which the mathematics goals of the lesson were aligned with standards-aligned content for the grade level taught	0 = not appropriate for grade level; 1 = appropriate
Discussion of student thinking (multiple solution paths)	The extent to which coach and teachers have deep and specific discussions of the multiple solution paths students might use to solve the task	0 = no solution paths discussed; 1 = all solution paths discussed in superficial ways; 2 = one discussed in more than superficial ways; 3 = at least two discussed in more than superficial ways
Discussion of pedagogy: advancing questions	The extent to which coach and teacher identify ways that teachers can advance student thinking toward the mathematical goals of the lesson using questioning	0 = questions not named; 1 = 1 or more questions named but not discussed; 2 = discussed with depth (e.g., likely student answers and teacher subsequent responses)
Discussion of math content	The extent to which discussion of mathematical goals went beyond broad topics (e.g., fractions) to include specific math concepts or principles	0 = no goal named; 1 = broad topics named; 2 = gives math definition or procedure as goal; 3 = discuss student acquisition of the underlying meaning of a concept
Students' opportunities to engage in conceptual thinking composite (average of maintenance of cognitive demand and attention to student thinking):		
Maintenance of cognitive demand	The extent to which the teacher maintains the cognitive demand of the lesson from materials, to set up, and through enactment	2–8 scale: 1–4 for maintenance from written to setup and from setup to enactment (see Stein & Kaufman, 2010) for procedures
Attention to student thinking	The degree to which teachers explored and facilitated the public display of student thinking throughout the lesson	1 = the teacher did no work to uncover student thinking; 2 = the teacher did some work to uncover student thinking, including asking students to publicly share their work; 3 = in addition to #2, the teacher purposefully selected some students to share their work; 4 = in addition to #2 and #3, the teacher connected or sequenced students' responses in a meaningful way. Item converted to a 2–8 scale.
Coach assist	Degree to which the coach assisted the teacher while they were teaching the task	0 = no coach help; 1 = minor prompts from the coach to the teacher; and 2 = the coach, at times, helped coteach the lesson

the Year 2 baseline and postcoaching lessons submitted for each partner teacher. Our composite measure—Students' opportunities to engage in conceptual thinking—is an average of two measures: maintenance of cognitive demand and attention to student thinking. Our coaching framework focuses on building the capacity of teachers to orchestrate high level mathematics tasks in the classroom, because prior research suggests that cognitively demanding tasks provide opportunities for students to build conceptual understanding by engaging in productive struggle (Stein & Lane, 1996). The specific measures (maintenance of cognitive demand and attention to student thinking) assess teacher capacity to orchestrate such tasks and have been shown to be associated with student learning gains (Stein & Lane, 1996), in one of the main studies cited by Hiebert and Grouws's (2007) in support of their

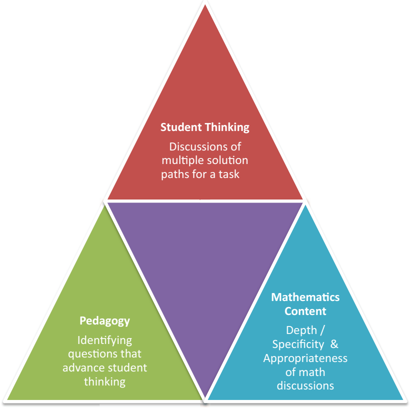
Student Thinking
Discussions of multiple solution paths for a task
Pedagogy
Identifying questions that advance student thinking
Mathematics Content
Depth / Specificity & Appropriateness of math discussions

FIGURE 2. *Dimensions of deep and specific discussions of the instructional triangle.*

claim that productive student struggle is associated with student learning.

Cognitive demand measures have been widely used to judge the extent to which mathematics teachers adhere to the main tenets of ambitious, standards-based instruction (e.g., Boston & Smith, 2009; Jackson et al., 2013; Rigby et al., 2017; Stigler & Hiebert, 2004). We employed the procedures described by Stein and colleagues (1996) to code cognitive demand as written in the task, at lesson set-up, and during lesson enactment. Following procedures described in Stein and Kaufman (2010), we calculated a "score" for the maintenance of cognitive demand throughout the lesson. Specifically, we measured the maintenance of cognitive demand from the task-as-written to task-as-setup (rubric score from 1 to 4) and from task-as-setup to task-as-enacted (rubric score from 1 to 4), and then summed them creating a scale from 2 to 8. In addition, raters scored the degree to which teachers explored and facilitated the public display of student thinking throughout the lesson on a scale from 1 to 4. We adjusted the scale of the latter item to develop a mean of the two scales—yielding a composite on a scale from 2 to 8. A higher score on the students' opportunities to engage in conceptual thinking composite therefore represents not only maintenance of cognitive demand of the task during the lesson, but also whether students had the opportunity to engage in and make public their (conceptual) thinking.

The videos were scored by a set of seven mathematics education experts, primarily assistant professors in universities who were trained to utilize the scoring rubric and subsequently scored 410 classroom videos. In Year 1, videos were scored after each cycle. In Year 2, the videos were scored only after all videos had been collected. In Year 2 scoring, raters were blind to both the teacher they were scoring and also the cycle the video was from. During the scoring of Year 1 videos, we randomly selected 33 of the 176 videos (almost 20%) for double-scoring. Videos were evenly distributed among our raters. We examined the interrater reliability of students' opportunities to engage in conceptual thinking, as this is the measure used in all subsequent analyses. Given our measure was scored from 2 to 8 and is ordinal, we examined the intraclass correlation coefficient (ICC) between two raters for our 33 double-scored videos. ICCs, equivalent to Cohen's weighted kappas, are commonly considered "good" when they are between .6 and .74, while ICCs above .74 are considered "excellent" (Hallgren, 2012). We obtained an ICC of .62 indicating adequate interrater reliability for our measure of students' opportunities to engage in conceptual thinking.

In addition to scoring that used the measures described above, raters also scored each lesson for the degree to which the coach assisted the teacher during implementation of the task. "Coach assists" were scored on a 3-point scale with "0" indicating no coach help, "1" indicating minor prompts from the coach to the teacher, and "2" indicating the coach, at times, helped coteach the lesson. With respect to interrater reliability, raters applied very similar scores on lessons where they agreed there were no coach assists (ICC = .72; n = 22) versus lessons where at least one of them coded coach assists (ICC = .52; n = 11).[2]

Table 2 provides an overview of the final data set for the study accounting for missing data and variability in the length of time that partner teachers participated in the study. Overall, we had limited missing data: we gathered and scored 97% of expected prelesson planning conferences and 93% of expected lesson videos. We also employed analytic models that enabled us to accommodate missingness.[3]

Analytic Approach

Our analyses sought to explore trends in coaching practice, trends in teaching practice,

TABLE 2
Sample of Partner Teachers and Data Availability

Partner teacher participation	*N*	Complete data	Teachers with one incomplete data cycle	Teachers with two cycles of incomplete data
Participated in Years 1 and 2 (seven data cycles)	24	18	5	1
Participated in Year 1 only (three data cycles)	41	36	5	0
Participated in Year 2 only (four data cycles)	38	19	15	4
Total teachers	103	73	25	5

and the relationship between coaching and teaching. Although these analyses are primarily descriptive, our longitudinal exploration of coaching and teaching aimed to generate suggestive evidence regarding the relationship between coaching and instruction.

Statistical Analyses for Growth in Coaching. To understand whether the depth and specificity of prelesson planning conferences improved over time, we explored trends in two ways. First, we examined mean scores and standard deviations for each item contributing to the measure of preconference depth and specificity for each cycle. A one-way repeated measures analysis of variance (ANOVA) was conducted to determine the effect of time on each item,[4,5] as well as the composite of the four indicators (the mean of the four scale-adjusted items). These analyses demonstrate whether there was a mean change, in general, over time, when including all preconferences across all teachers and coaches. The repeated measures ANOVAs also provide descriptive statistics about change in the coefficient of variation[6] over time, as well as measures of variance to estimate the standard deviation to calculate our within-subjects effect size estimates.[7]

Second, in an attempt to describe overall patterns of growth and develop estimates of the magnitude of this growth, we examined hierarchical linear growth models. In these models, time points are nested within coach–teacher pairs, to better understand patterns of within-subjects change over time. Using HLM v.7.03, we examined five separate univariate analyses[8]—one for each item measuring depth and specificity of the preconferences and one for the composite score. To investigate whether different items grew at different rates during Year 1 and Year 2, we examined piecewise hierarchical linear growth models (Raudenbush & Bryk, 2002, pp. 178–182). Our two-level unconditional model for the composite can be summarized as follows (Appendix A1, available in the online version of this journal)[9,10]:

Level 1:

$$\begin{aligned}\mathrm{Pre-Conf.}\mathit{Depth}_{ti} &= \pi_{0i}\\ &+ \pi_{1i}\left(\mathrm{Base\ Rate}_{ti}\right)\\ &+ \pi_{2i}\left(\mathrm{Increment}_{ti}\right) + e_{ti}\end{aligned} \tag{1.1}$$

Level 2:

$$\begin{aligned}\pi_{0i} &= \beta_{00} + r_{0i}\\ \pi_{1i} &= \beta_{10} + r_{1i}\\ \pi_{2i} &= \beta_{20} + r_{2i}\end{aligned} \tag{1.2}$$

Our primary interest in these models was to understand and describe model-based rates of growth over time for depth and specificity of prelesson planning conferences for coaches trained during the first 2 years of development of the TN + IFL Coaching Model. Model-based estimates allow us to construct the average growth trajectory using all data from teachers who experienced coaching during the intervention (i.e., 311 preconferences with 103 teachers, $n_j = 3.02$). We were also able to determine the statistical significance of each interval in these growth trajectories, whether some items displayed greater increases during particular intervals, and the variability between teachers in their growth rates for each interval.

Statistical Analyses for Growth in Teaching. We examined a series of HLM models to answer different questions with our teaching data. Using all videoed observations across both study years we examined model-based estimates of change in teaching among our treated sample. The purpose of these analyses was to explore teaching growth in the context of a coaching intervention. The final analyses incorporate all data from all time points (i.e., 410 videos from 103 teachers, $n_j = 3.98$).[11]

Our findings demonstrate what we learned as we engaged in a process of model-building within each of the different ways we configured our data to determine the best model fit.[12] In so doing, we examined whether the functional form of improvement for our coached teachers was linear, quadratic or cubic. We began the process by adjusting for several independent covariates. For example, because teaching scores could be affected by help from coaches we adjusted for whether raters considered the coach to have assisted during the enactment of the lesson.[13] Coach assistance during lessons also created ambiguity about how to score the teachers' performance for a given lesson, so we accounted for differences in how raters scored videos by including a fixed effect for rater as a dichotomous time-varying covariate. Finally, we also included a dichotomous time-varying covariate for whether the lesson video was obtained during a coached session or not, as videos during Year 2 included two un-coached videos—one at the beginning of the year (Year 2 baseline lesson) and one at the end of the year (Year 2 postcoaching lesson).[14] We ran a model to estimate the average growth in teaching across all teachers before examining prediction models with fixed effects at the teacher level of the model.

Examining associations between preconference depth and specificity scores and changes in teaching. To explore between-teacher differences in their growth trajectories, we also included a measure of the partner teachers' average preconference depth and specificity scores across cycles for coach–teacher pairs.[15] This model also adjusts for the teachers' beginning status in providing students' opportunities to engage in conceptual thinking. Below we present the prediction model from our final data configuration. To simplify the presentation of the model, we did not include the fixed-effect estimates of the Level-1 dichotomous time-varying covariates described above although all of those adjustments were included in our final models and results of those models are supplied in Table B3 in online Appendix B.

Level-1:

$$
\begin{aligned}
&\text{Opportunities for Students'}\\
&\text{Conceptual Thinking}_{ti} = \pi_{0i}\\
&\qquad + \pi_{1i} * (\text{Time}_{ti})\\
&\qquad + \pi_{2i} * (\text{Time Squared}_{ti})\\
&\qquad + \pi_{3i} * \text{Time Cubed}_{ti}) + e_{ti}
\end{aligned}
\tag{2.1}
$$

Level-2:

$$
\begin{aligned}
&\pi_{0i} = \beta_{00} + \beta_{01} * (\text{Init. Tch. at Ceiling}_i)\\
&+\beta_{02} * (\text{Avg. Pre-conference Depth}_i) + r_{0i}\\
&\pi_{1i} = \beta_{10} + \beta_{11} * \left(\text{Init. Tch. at Ceiling}_i\right)\\
&+\beta_{12} * \left(\text{Avg. Pre-conference Depth}_i\right) + r_{1i}\\
&\pi_{2i} = \beta_{20} + \beta_{21} * (\text{Init. Tch. at Ceiling}_i)\\
&+\beta_{22} * (\text{Avg. Pre-conference Depth}_i) + r_{2i}\\
&\pi_{3i} = \beta_{30} + \beta_{31} * (\text{Init. Tch. at Ceiling}_i)\\
&+\beta_{32} * (\text{Avg. Pre-conference Depth}_i) + r_{3i}
\end{aligned}
\tag{2.2}
$$

While all parameters of the model are described in the online files in Appendix A2, we briefly describe the decision to account for a teacher's beginning status. To adjust for beginning status we included a dichotomous indicator for whether the teacher was at ceiling on the maintenance of cognitive demand measure (a score of 8) on their first recorded lesson (with teachers at ceiling coded as 1). For teachers beginning in Year 1, the first recorded lesson was the one that occurred in the first coaching cycle; for teachers beginning in Year 2, it was the un-coached Year 2 baseline lesson. Fifteen percent of teachers were at ceiling on their first recorded lesson. We found this adjustment crucial because teachers beginning at or near the ceiling had no opportunity to demonstrate growth on measures of their teaching, but they did have potential for growth in preconference depth and specificity.[16]

Examining how within-teacher changes in preconference depth and specificity influence changes in teaching. Finally, we also ran one subsequent model beyond the one just described with the only change being the addition of preconference depth and specificity scores at Level

TABLE 3

Data Cycles

Data cycles	Description of data collected and analyzed in this manuscript
1	Coaching Cycle 1: Prelesson planning conference and lesson
2	Coaching Cycle 2: Prelesson planning conference and lesson
3	Coaching Cycle 3: Prelesson planning conference and lesson
4	Year 2 baseline lesson
5	Coaching Cycle 4: Prelesson planning conference and lesson
6	Coaching Cycle 5: Prelesson planning conference and lesson
7	Year 2 postcoaching lesson

1, person-mean centered, as a time-varying covariate. By centering preconference depth and specificity scores within teachers, this analysis examined, for each coached cycle, whether the deviation in preconference scores (from an individual's average preconference score) at each timepoint was predictive of their teaching practice. This model further helps us understand whether within-teacher changes in preconference depth and specificity scores influenced growth in their teaching practice, which can be considered a causal estimate when adjusting for between-teacher differences in preconference scores (see, for example, Duckworth et al., 2010). Here, again, we adjusted for whether the initial estimate of teaching practice contributed by a teacher was at ceiling because estimates of this relationship should be examined among those teachers with potential for both measures to increase over time. Throughout discussion of the findings we refer to data cycles (1−7), which correspond with data collection events including five coaching cycles and the two un-coached recorded lessons collected in Year 2 (see Table 3).

Findings

We developed a model for mathematics coaching practice, trained coaches to enact the model, and tested it in schools by collecting and analyzing data on its effectiveness. Our analyses provide promising evidence in support of key features of the model. Participating coaches did in fact use the model's key coaching practices during the two school years, as evidenced by analyses of videotaped coaching interactions. Likewise, partner teachers improved their capacity to provide rich opportunities for students to develop understandings of key mathematical concepts. Finally, we demonstrate how the depth and specificity of coaching conversations predicted rates of improvement in teaching practice, and how within teacher changes in the depth and specificity of conversations also predicted growth in teaching.

RQ1: To What Extent and How Did Coaching Improve Over Time?

Growth in Coaching Practice. Each of the four items measuring deep and specific prelesson planning conversations demonstrated statistically significant improvement over time. Raw means, standard deviations, and the coefficient of variation for each cycle are all reported in Table 4. For each item, there is an increase in the mean, as well as a decrease in the standard deviation, resulting in a lower coefficient of variation (dispersion relative to the mean) over time. There was a marginally significant effect of time on: (a) *appropriateness of mathematics content*, Wald $\chi^2 = 8.95$ ($df = 4$, $p = .06$); and a statistically significant effect of time on (b) *discussion of student thinking: multiple solution paths* $F(4, 247.66) = 4.40$, $p = .002$; (c) *discussion of pedagogy: advancing questions,* $F(4, 249.00) = 4.96$, $p = .001$; (d) *discussion of specific mathematics content,* $F(4, 248.84) = 9.54$, $p = .000$; and (e) *composite* pre*conference depth and specificity,* $F(4, 249.66) = 6.04$, $p = .000$.

To examine non-linear trends for some items suggested by the raw mean values in Table 4, we examined piecewise hierarchical linear models with our data. We first examined whether the piecewise model was a better fit to the data for each of the five univariate analyses. Three of the

TABLE 4

Repeated Measures Descriptive Statistics for Depth and Specificity of Prelesson Planning Conferences at Each Cycle

	Appropriateness of math content for grade level[a]			Discussion of student thinking: Multiple solution paths		Discussion of pedagogy: advancing questions		Discussion of specific math content		Composite preconference depth and specificity	
Data cycle	Number of PC	*M* (*SD*)	COV (%)	*M* (*SD*)	COV (%)	*M* (*SD*)	COV (%)	*M* (*SD*)	COV (%)	*M* (*SD*)	COV (%)
1	63	0.83 (0.38)	46.3	1.60 (1.01)	63.3	1.55 (1.14)	72.5	2.13 (0.73)	34.0	1.94 (0.69)	35.2
2	65	0.88 (0.33)	37.9	1.85 (0.87)	47.1	1.82 (1.04)	57.2	2.09 (0.86)	41.2	2.10 (0.61)	29.1
3	64	0.86 (0.32)	37.2	2.06 (0.81)	39.5	1.48 (1.12)	75.7	2.69 (0.50)	18.6	2.20 (0.55)	24.8
5	61	0.95 (0.20)	21.3	1.98 (0.94)	47.4	1.99 (1.12)	56.2	2.49 (0.67)	27.0	2.33 (0.59)	25.4
6	58	0.95 (0.23)	24.2	2.22 (0.82)	36.8	2.17 (1.06)	48.7	2.43 (0.70)	28.9	2.42 (0.54)	22.4

Note. PC = prelesson conferences; *SD* = standard deviation; COV = coefficient of variation.

[a] Estimated marginal mean values and dispersion statistics were derived from a repeated measures logistic regression model.

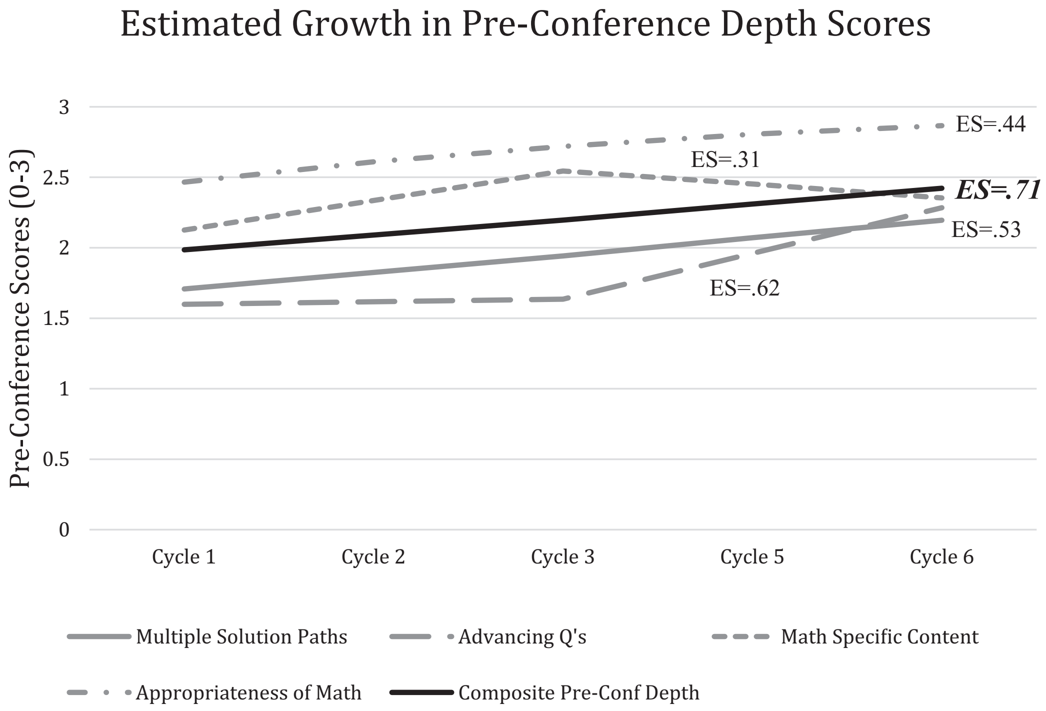

FIGURE 3. *Plot of growth estimates of coaching from the piecewise hierarchical linear models.*

analyses demonstrated a better model fit as a linear model (i.e., the test for an increment beyond the base rate by itself did not result in a significant reduction in deviance). In each case, the linear base rate demonstrated significant growth in depth and specificity across the 103 coach–teacher pairs—*appropriateness of mathematics content* (β_{10} = 1.19 logits/year; p = .021), *discussion of student thinking: multiple solution paths* (β_{10} = .356 points/year; p = .010), and *composite* pre*conference depth and specificity* (β_{10} = .319 points/year; p = .001). For model parsimony, for each of these three items we retained the growth estimates from the linear base-rate-only model for Figure 3. For two of the items, *discussion of pedagogy: advancing questions* (χ^2 = 6.15, df = 3, p = .100) and *discussion of specific math content* (χ^2 = 11.70, df = 3, p = .009) the piecewise models identified a significant increment beyond the base rate between the end of Year 1 (third round of coaching, Data Cycle 3) and beginning of Year 2 (fourth round of coaching, Data Cycle 5). These models thus provide two different linear growth rates for Year 1 and Year 2 for these two items, as reflected in Figure 3.

Effect sizes shown in Figure 3, provide a sense of the relative magnitude of these findings. We generated within-subject effect sizes (d_{av}) based on the total improvement seen in the model-based estimates, relative to the pooled standard deviation in the outcome. Model-based estimates represent improvement over a year-and-a half, given the first round of coaching was in the fall of AY 2015–2016 (Data Cycle 1) and that the fifth round of coaching (Data Cycle 6) happened in the early spring of AY 2016–2017. As displayed in Figure 3, effect sizes for individual items ranged from d_{av} = .31 to d_{av} = .62. The effect size for the composite score d_{av} = .71 indicates that, overall, coach–teacher pairs were growing in preconference depth and specificity over time, and that, on average coach–teacher pairs improved about seven tenths of a standard deviation of the outcome.

More substantively, with items converted to be on the same 0 to 3 scale, we see that by the fifth round of coaching (Data Cycle 6), each individual item, on average, is between a score of "2" and "3"—the top two rubric scores, indicating that not only did coach–teacher preconferences grow in their depth and specificity, but they are far closer to the top of the scale than the bottom (for the meaning of the scale see "Metric" column in Table 1). Indeed, the average for the composite depth and specificity score is approaching 2.5. The growth coefficient (β_{10} = .338 points/year; p = .001) describes the average gain in preconference depth and specificity per year. Thus, teachers grew nearly one half point, on average, over the five rounds of coaching, representing a shift, generally, from pro-forma implementation of multiple solution strategies or discussion of superficial elements of the task and toward a deeper consideration of implementing multiple strategies and/or discussion of students' acquisition of the meaning of concepts.

Growth in Coaching Practice Aligned With Training Program. Figure 3 shows a graphical depiction of the overall improvement in coaching conversations, in general, where the items are all adjusted to the same 0 to 3 scale. The non-linear (and inverse) patterns for two of the items (deep and specific discussions of *advancing questions* and *math content*) are evident in this visual display of the growth trajectories for coaching practice. Across all coach–teacher pairs, coaching conversations were seen to first improve in having greater depth and specificity of discussions of the math content for the lesson, and in subsequent cycles showed more rapid growth in their depth of discussions about advancing questions. As we considered these patterns in relation to our projects improvement cycles, we see that these trends in growth trajectories align with the focus of our coach training across network meetings, lending further support for our hypothesis that training coaches to have deep and specific conversations with teachers would influence their capacity to do so in prelesson planning conferences.

For example, in the network meeting immediately prior to the third round of coaching (March 2015), we emphasized the need to develop students' mathematical understanding and engaged coaches in activities that pressed them to identify, "what mathematics do we want students to know or understand as a result of implementing this task" as they analyzed mathematics tasks. In addition, coaches were trained in the difference between performance mathematics goals and learning goals. For example, they were given examples of how identifying learning goals such as "students will understand and recognize that a unit rate describes how many units of the first quantity correspond to one unit of the second quantity" provide greater guidance for teachers as they teach for conceptual understanding than a performance goal such as "students will calculate the unit rate by determining the ratio to one." It is not surprising then that the prelesson planning conferences following this meeting included more in-depth discussions of the specific mathematical content for the lessons coaches and teachers were collaboratively planning.

Subsequently, our analyses of the prelesson conferences from the first year of the project, led us to identify the need to provide additional training on pedagogy to support the development of students' conceptual understanding of mathematics. In the fourth network meeting (August 2015), coaches had opportunities to identify how they could guide prelesson planning conferences to discuss in-depth pedagogy for supporting student learning. Specifically, coaches analyzed transcripts from prelesson conferences for evidence of deep and explicit discussion of the instructional triangle. In particular, we emphasized discussion of pedagogy, including explicit attention to how teachers can plan questions to advance student understanding. It is not surprising then that we see more in-depth discussion of advancing questions in the fourth and fifth rounds of coaching (Data Cycles 5 and 6) following this meeting. These patterns may suggest that given limited time in the prelesson planning conferences, coaches made choices to emphasize discrete aspects of this practice at different times as they were learning how to incorporate deep and specific conversations into their practice.

Coaching Growth Trajectories Do Not Vary Significantly. One important question for gauging our effectiveness was the degree to which all coaches (and their associated coach–teacher pairs) seemed to have benefited from the training. In other words, were effects obtained because we observed improvements among a handful of coaches and their associated teachers,

TABLE 5

Repeated Measures Descriptive Statistics for Video-Based Measure of Opportunities for Students to Engage in Conceptual Thinking During Lesson at Each Cycle

			Opp. for students' conceptual thinking			Coach assists	
Cycle	Coached cycle?	Number of videos	*M*	*SD*	COV (%)	*M*	*SD*
1	Yes	62	5.57	1.85	33.1	0.63	0.77
2	Yes	65	5.85	1.61	27.6	0.78	0.86
3	Yes	62	5.81	1.64	28.2	0.41	0.75
4	No	60	5.01	1.82	36.4	0.10	0.30
5	Yes	55	5.67	1.45	25.4	0.45	0.63
6	Yes	53	6.35	1.27	19.9	0.38	0.60
7	No	53	6.49	1.23	18.9	0.23	0.42

Note. SD = standard deviation; COV = coefficient of variation.

or were improvements seen broadly across most coach–teacher pairs? Variance decompositions from the HLM analyses show there is *not* significant variation between coach–teacher pairs on the growth slope ($\tau_{1;}$ $\chi^2 = 96.98$; df = 98, $p > .500$—see Table B1 in online Appendix B). Other features of this model are also important to mention. For example, when examining a linear model for the composite depth and specificity of preconference conversations, about half of the variance was between time points within coach–teacher pairs (i.e., measurement error: σ^2). It is notable that this model demonstrates high variability between preconference scores within teachers (σ^2) which serves to underscore one of the difficulties of measuring change in the quality of coach–teacher conversations over time. In addition, there was significant variance between coach–teacher pairs in their baseline preconference depth and specificity at Cycle 1 ($\tau_{0;}$ $\chi^2 = 130.33$, $df = 98$, $p = .016$). Thus, in this rather idealistic implementation of a mathematics coaching model (i.e., relative lack of resource constraints limiting coaches from conducting full coaching cycles with two partner teachers), coach–teacher pairs significantly differed at baseline (τ_0), but they did not significantly differ from the average rate of improvement (τ_1). There is further evidence that coaches also did not vary significantly in the growth rates experienced by their two partner teachers relative to the rest of the group. When we ran three-level models, where coach–teacher pairs were nested in coaches, the growth rates between coaches also did not significantly vary ($\tau_{\beta 1;}$ $\chi^2 = 34.84$; $df = 32$, $p > .500$—see Table B2 in online Appendix B), nor did the growth rates between-teachers within-coaches ($\tau_{\pi 1;}$ $\chi^2 = 74.81$; df = 66, $p > .500$—see Table B2 in online Appendix B). Given this, it is notable that the variance components show significant differences between coach–teacher pairs in their status but not in their average growth rate.

RQ2: To What Extent Did Teaching Improve Over Time?

The findings from models examining changes in teaching largely parallel the findings for growth in the depth and specificity of preconference conversations. Teaching improved for almost all partner teachers. Arriving at accurate effect-size estimates to describe the patterns of teaching improvement, however, is complicated by several factors including, the presence of coach assists, the fact that two of the seven videos were un-coached and occur in Year 2 of the study, that participants vary in how long they were coached, and that ceiling effects (even at the beginning of coaching) were present for some of our partner teachers. A quick examination of the raw data, provided in Table 5, is illustrative of both the underlying growth and these complications.

A glance at the progression of means over cycles shows that, generally, scores are higher in Cycle 7 (an un-coached cycle with virtually no coach assists) than they were in Cycle 1, supporting the notion that growth in teaching was occurring in the population of partner teachers, in general. In addition, the standard deviation is decreasing over cycles and so is the coefficient of variation. Yet, while the number of videos remains consistent across cycles, different teachers contribute to the means, hence the need for hierarchical growth models to more appropriately estimate within-teacher changes over time. In addition, the lack of growth, but not decline, in the mean for Cycle 3 of Year 1, masks potential improvement represented in the reduction in coach assistance during teaching events from Cycle 2 to Cycle 3. If coach assists diminish, but the teaching scores remain the same, then this represents "growth" because teachers achieve the same score with less teaching assistance from their coach (i.e., their scores are obtained independent of their coach). Declining coach involvement during lessons likely represents an important improvement in the coaching process by placing more responsibility onto the teachers for instruction. Indeed, coach assists demonstrate a continued decrease over the cycles.[17]

Further evidence for teaching improvement is shown in the teaching growth trajectory presented in Figure 4. This graph represents the model based estimates of average growth in teaching during the approximately 1½ year interval we were training coaches. On average, teachers gained about 1.51 points on our measure of teaching across this time interval. The effect size of .95 is achieved despite the fact that 31% of teachers started between 6 and 8 on our scale that ranges from 2 to 8, and an additional 15% of teachers started at 8. Thus, there was a high proportion of teachers near or at the ceiling on our scale. Finally, our model estimates suggest the per-year effect size is .61. The online Appendix A3 provides further details about how this finding on a per-year basis is equivalent to alternate data configurations (see Table A3.1.). Furthermore, in comparing and contrasting across models, we found that, in general, all teachers contribute to the overall growth estimate (i.e., we did not see differences in growth for first year only teachers vs. second year only teachers, and teachers present for both years continued to grow in their second year of the study).

RQ3: Does Variance in the Depth and Specificity of Teachers' Prelesson Planning Conversations Predict Growth in Teaching?

To better understand the association between the depth and specificity of preconference conversations and growth in teaching, we added the composite measure preconference depth and specificity as a fixed effect to the model (see left hand column of Table B3 in online Appendix B). Figure 5 illustrates the association between preconference depth and specificity and the rate of growth in teaching. This graph shows different trajectories based on average preconference depth and specificity scores. The solid line represents 23 teachers whose preconference depth and specificity scores were .5 *SD* or more below the mean; ES = .66 for growth in teaching over the 1½ year interval. The dashed line represents 54 teachers whose depth scores were between −.5 *SD* and .5 *SD*; ES = .95 for growth in teaching. And the dotted line represents 26 teachers whose depth scores were .5 *SD* or more above the mean; ES = 1.26. In other words, teachers whose coaching conversations were characterized by greater depth and specificity in prelesson planning conferences had higher rates of growth in teaching.

Examining Whether Changes in Depth and Specificity of Coaching Conversations Predicts Change in Teaching. To better understand the association between changes in coaching conversations and changes in teaching, we added a within-person measure of preconference depth and specificity as a fixed effect to the model.[18] In addition to the between teacher differences in growth trajectories described above, the time-varying covariate for within-teacher change in standardized composite preconference depth and specificity scores was also marginally statistically significant (β_{120b} = .20, $p = .055$; see right hand column of Table B3 in online Appendix B). Thus, in addition to the between-teacher differences in growth noted earlier, for each one standard deviation change in preconference depth and specificity scores within-teachers, they are predicted to gain an additional .20 points in their teaching, ultimately providing students' greater opportunities to engage in conceptual thinking during mathematics lessons. These within-person effects show that growth in depth and specificity of preconference

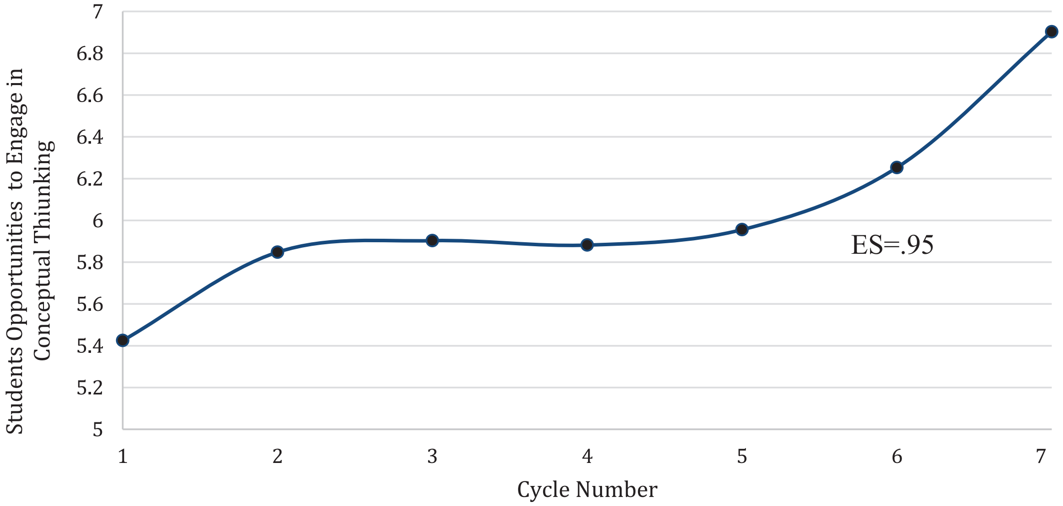

FIGURE 4. *Two-year model-based cubic growth trajectories in classroom teaching scores for all 103 partner teachers, with grand-mean centered adjustments for coach assists, video rater, and coached versus un-coached lessons.*

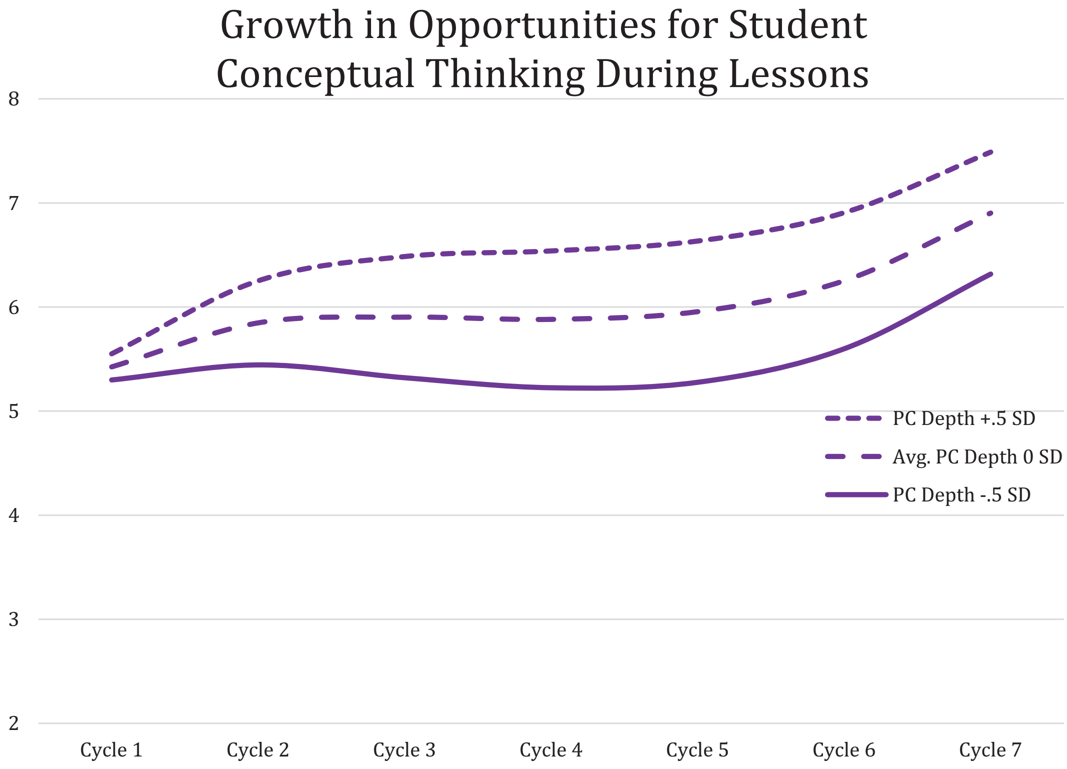

FIGURE 5. *Comparing model-based estimates of two-year cubic growth trajectories for all 103 partner teachers at different levels of average preconference depth and specificity scores, with grand-mean centered adjustments for coach assists, video rater, and coached versus un-coached lessons.*

conversations predicts changes in teachers' providing students' opportunities to engage in conceptual thinking during lessons.

Discussion

Our findings suggest that when coaches had deep and specific conversations with teachers in the context of planning specific lessons—including attention to content, pedagogy, and student learning—teachers improved their capacity to maintain the cognitive demand of high-level mathematics tasks. Developing this capacity is significant because prior research suggests (a) that maintaining the demand of high-level tasks is challenging (Stein et al., 1996, 2010) and (b) that when teachers provide opportunities to reason through complex tasks and sustain student engagement at a high-level, students are more likely to learn mathematics concepts (Boaler & Staples, 2008; Hiebert & Wearne, 1993; Stein et al., 2017; Stein & Lane, 1996; Stigler & Hiebert, 2004; Tarr et al., 2008). Teaching that supports students' conceptual understanding of mathematics is critical in the current policy environment that has set rigorous college- and career-readiness standards, as the learning goal for all students.

Through our ongoing analyses of data from successive improvement cycles, we came to understand how prelesson planning conferences, and specifically deep and specific discussions of the instructional triangle, present a critical opportunity for teachers to prepare for instruction. By tying discussion of mathematics lessons to important dimensions of the instructional triangle, teachers have a template for learning how to incorporate broad ideas about conceptually oriented and student-focused instruction into their practice. In addition, our results suggest that these coaching practices were associated with growth in teaching even when teachers had as little as two or three coaching cycles in a given year. Orchestrating deep and specific prelesson planning conferences appears to be a high-leverage coaching practice.

This study makes a significant contribution to prior research on coaching by examining coaching practice in a large sample of coach and teacher discussions. The majority of coaching studies have treated coaching practice as a black box or investigated it in small numbers of qualitative cases (Kraft et al., 2018). By identifying a high-leverage coaching practice and a way to measure its uptake in practice, we provide a model for the kind of research that is critical to advance the field's understanding of coaching. In addition, the findings make a tentative connection between the quality of the uptake of specific coaching practices and growth in teaching practice.

Although the analyses presented in this article provide significant insight into high-leverage coaching practice during prelesson planning conferences, there are limitations to our analyses that should be considered when interpreting our results and which provide guidance on avenues for future research. First, the way we measured instruction—operationalized in the context of or close proximity to coaching cycles—is not a measure of the extent to which teachers changed their typical mathematics teaching. Rather this is a measure that provides an indicator that teachers have developed capacity to maintain the cognitive demand of high level tasks. This is an important first step in establishing the potential power of the coaching model, yet the lack of attention to sustained change in practice is a potential limitation that creates an opportunity for future research.

Similarly, it may be possible that the changes in teaching we observed in our sample could be caused by factors other than exposure to our coaching model. For example, teaching practice may be improving due to increased exposure to or understanding of the rigorous college and career standards that occurred outside the context of the coaching study. Our longitudinal growth models provide suggestive evidence that coaching practice is associated with the changes in teaching practice that we observed because within teacher changes in the depth and specificity of coaching conversations predicted growth in their teaching. However, a similar concern arises in that these changes could be attributed to other exogenous factors such as teachers' interests in developing student conceptual understanding, which may have facilitated both growth in the depth of prelesson planning conferences and teaching for understanding. This suggests that further research is necessary to substantiate the relationship between this coaching practice and teacher learning and development.

Furthermore, a rival hypothesis for the changes in instruction we observed may simply be that teachers were encouraged to utilize a challenging task. In other words, the coaching effects may stem primarily from the selection and faithful implementation of high-quality tasks rather than the quality of the coaching itself. Although possible, we argue that this it is not likely that coaching effects can be traced simply to the selection of tasks, because implementation of high-level tasks is hard to do without support for learning the pedagogical strategies that support maintenance of cognitive demand (Stein et al., 1996).

Another limitation of our study is that we relied on a carefully selected group of coaches, who in turn selected participating teachers. As a result, we do not know the extent to which the findings might generalize to more typical coaching contexts. Although we took care to select coaches that varied in their baseline capacity and prior training, the sample may not reflect the full range of coaches in natural contexts. In addition, the range of teacher skills and capacity may not be representative. We acknowledge this as a limitation and note that future investigations should examine whether typical coaches and teachers achieve similar gains when they are trained to utilize the coaching and teaching practices described in this study.

Our research contributes significantly to future explorations that can begin to disentangle the influence of exogenous factors on teaching growth. Having identified a seemingly high-leverage coaching practice, we create opportunity to test whether coaches trained to enact this practice produce superior gains in teaching development than teachers coached by coaches without that training. In a follow up investigation we are analyzing results from a prospectively matched quasi-experimental study that compares coaching and teaching effects for two different groups of teachers—those coached using our model versus those receiving garden-variety coaching.

Our findings have significant implications for research, policy, and practice. In addition to the contributions to research on coaching noted above, our study illustrates the affordance of continuous improvement research conducted in the context of a research–practice partnership. By partnering with coaches, we had an opportunity to collect rich and comprehensive data on coaching practice and gain insights into complex implementation dynamics. In this way, our work is part of a broader trend in education research that aims to utilize research–practice partnerships to facilitate systemic educational improvement and knowledge production (Coburn & Penuel, 2016). We believe our experience reinforces the importance of research–practice partnerships, providing an example of how researchers, policymakers, and practitioners can work together to support ambitious instructional improvement. The study exemplifies how a design-based approach to conducting implementation research can result in improved policies and practices while also generating research findings that are useful to the field.

With respect to policy and practice, our study supports schools and districts aiming to utilize instructional coaching as part of their improvement agendas. With the growing investment in instructional coaching in districts around the country that are trying to support shifts in teaching aligned with rigorous standards, a need has emerged for providing guidance and training for the coaching role (Gallucci et al., 2010; Kraft et al., 2018; Matsumura et al., 2009). To inform the design of coaching programs, we need rigorous empirical examinations of what coaching practices contribute to teaching improvement, so schools and districts can get an optimal return on this investment. Our investigation identifies a seemingly high-leverage practice that our experience suggests can be taught to coaches, and productively incorporated into their practice. The Coach–Teacher Discussion Process routine with a focus on deep and specific prelesson planning conferences provides considerable guidance for how coaches can utilize their time in support of teacher learning and practice improvement. In addition, it suggests a focus for coach training, support, and evaluation.

This kind of practical knowledge about coaching is critical as local policymakers and instructional leaders implement ambitious instructional improvement efforts that aim to ensure equitable access to the learning opportunities students need to achieve college and career readiness standards. Coaching is a component of the instructional guidance infrastructure that districts and schools can design and leverage to promote teacher learning and development (Cobb et al., 2018; Hopkins et al., 2013). District work to redesign the instructional guidance infrastructure can support teacher leadership and act as a coupling mechanism that ties district-level instructional priorities to teachers and their instruction (Hopkins & Woulfin, 2015). In addition, instructional coaching can support implementation of instructional reforms by working in concert with other reforms, such as teacher evaluation systems (Woulfin & Rigby, 2017). Studies that generate insights about educational practice, in this case the work of coaches to support teaching improvement, can provide the guidance necessary to help systems move beyond the identification of structures and policies that signal alignment with ambitious teaching and learning, to the design of robust instructional infrastructures that support teaching, learning, and continuous improvement.

Declaration of Conflicting Interests

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

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305H140112 to University of Pittsburgh. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Notes

1. As three of the four items were on a scale from 0 to 3, we transformed the last item to the same scale before taking the average of the four items.

2. Although this is a small sample to make inferences from, it points to one potential difficulty in scoring coached videos, that is, how to handle the scoring of lessons when the coach actively contributes to aid in the implementation of the lesson. These ICCs suggest that raters agree quite well when all of the teaching during the lesson is attributed solely to the teacher. However, there are possible rater by coach assist interactions which could lead to greater variability in rater's scores of teaching on coach-assisted lessons. Interrater agreement was lower for lessons where at least one of the raters indicated a coach assist. In our view this may be due to the complexity of scoring video-based lessons in the context of coaching. Should teachers be credited with the teaching achieved (ignoring that the idea may have originated with the coach) or should they only be scored on what the teacher did independent of the coach? This is a complex judgment raters face when conducting scoring and our hypothesis is that one reason the ICC might be lower for videos with assists is that different raters might have different views about how they should score these situations. In generating model-based estimates of teaching growth, we have adjusted for any main effects of a lesson being labeled as a "1" or "2" for coach assist, as well as any main effects of raters, in part to account for any such differences.

3. We used whatever data teachers had. One of the advantages of hierarchical models (and HLM software, in particular) is that it allows for missing data at Level 1. Therefore, it will construct within person growth estimates based on the portion of data a person has available.

4. We used a linear mixed model in SPSS v.26 and compared a compound symmetry model (constant variance assumed) versus an unstructured model (independent variance). Both models demonstrate a significant effect of time ($F = 6.84$; $p < .001$ vs. $F = 5.93$; $p < .000$). There is not a significant drop in −2 log likelihood for the unstructured model χ^2 (25, $n = 103$) = 22.80, $p = .59$. The more parsimonious model (compound symmetry) is preferred, and, thus, is the one we present in this article, though both models demonstrate a significant effect of time.

5. One item, the grade-level appropriateness of the mathematics content, was dichotomous and violated the assumption of normality for a repeated measures ANOVA outcome. To assess the statistical significance for change over time for this item we examined a repeated measures logistic regression with cycle as the focal independent variable.

6. The coefficient of variation is a dispersion statistic that measures the standard deviation relative to the mean. In studies with interventions it is likely that in addition to mean changes, the standard deviation might also decrease if subjects at the lower end of the distribution are "pulled up." The coefficient of variation, expressed as a ratio, would demonstrate a decrease if there was either an increase in the mean or a decrease in standard deviation, and would be especially large if both were occurring simultaneously.

7. For effect sizes we report Cohen's *d* for within-groups designs (d_{av}) as discussed in Lakens (2013). To find the average standard deviation across all time points, we used the standard deviation for the first and last time points from the descriptive statistics of the repeated measures mixed model and averaged them.

8. For this analysis we used 311 preconference videos from 103 teachers ($n_j = 3.02$) across both years. We examined two different types of growth models—we examined both two- and three-level univariate growth models. In the three-level models, in addition to time points nested in teachers, teachers were also nested in coaches ($n_j = 3.18$). We chose to present the simpler two-level growth models, given the fact that both produced essentially the same growth coefficients, that there was no significant variation between coaches in the growth estimates, and because the number of teachers nested in coaches is relatively sparse. In addition, this fits with our primary purpose which was to describe the observed growth parameters among all teachers.

9. Model parameters are described in online Appendix A1.

10. The "Base Rate" measures linear growth over the entire interval, while the "Incremental Rate" provides a significance test to identify whether the linear growth rate is the same in the second year as it was in the first year. If there is no change to the rate of growth in Year 2, then the "Incremental Rate" would be close to "0," indicating no divergence in growth rate.

11. Prior to examining our final data configuration, we examined two simpler growth models. The first model estimated teaching change over a single year across all 103 teachers. The second model estimated teaching change over 2 years for the 24 teachers receiving coaching in both years. The models and findings are described in online Appendix A3. In particular, Appendix A3 demonstrates the similarity in the magnitude of the effect size estimates on a per-year basis across the different model configurations. We justify our final model selection, including adjustments for covariates, because it uses all of our collected data.

12. The process of determining model fit included an examination of a chi-square test of the deviance statistics from two models, one of which was fully nested in the other. If the model with the greater number of parameters resulted in a reduction in deviance, taking into account the number of extra parameters added to the model then the model was deemed to be a better fit to the data.

13. Recall that coach assists were scored at three different levels. We created two dichotomous variables to be included as time-varying covariates in the model. The first dichotomous variable was for a coach assist of "1" (CchAssist1) when coaches provided suggestions to teachers, and the second was for a coach assist of "2" (CchAssist2) when coaches may have helped to coteach the lesson at times. Both dichotomous variables were entered in the model and videos with no assists were the reference category.

14. Inclusion of this variable muted the effect of coach assist one because coach assists were close to zero during these un-coached sessions. Thus, prior to this covariate being added the effects of coach assists appeared much greater, partly because it was picking up on the contrast of scores within teacher during these un-coached sessions. In addition to fewer assists, this covariate accounts for the fact that teachers also do not have preconferences for these two videos.

15. Although this measure is endogenous to their teaching gains, our intent here is merely to describe differences in these growth trajectories, and not make a causal attribution.

16. We examined many different models making adjustments for baseline level of teaching, including linear adjustments, to understand the sensitivity of model estimates to the model specification. In general, the findings persist no matter how adjustments for prior teaching status were made. Linear adjustments made less sense to us theoretically because we assumed the relationship might be curvilinear. We also tried grouping teachers into three groups—those 1 *SD* or more above in baseline status, those between 1 *SD* and −1 *SD* and those below −1 *SD* in status at baseline—with similar findings on the coefficients. Although lower standard errors for coefficients from these models indicated greater precision, an increase in variance components for the final model in Table B3 indicated the model may have over-controlled for beginning status. Therefore, we ended up only adjusting for whether or not the teacher was at ceiling at baseline.

17. In a repeated measures ANOVA, there was a statistically significant effect of cycle on coach assists $F(6, 336.32) = 13.35$, $p = .000$, demonstrating a decrease over time.

18. To run this model, we had to remove the time-varying covariate for coached versus un-coached lessons because un-coached lessons do not have preconference rigor scores. Thus, this model is primarily informative about the associations between coaching and teaching. We tested for a random effect but deviance statistics suggested this model was not a better fit.

References

Allen, J. P., Pianta, R. C., Gregory, A., Mikami, A. Y., & Lun, J. (2011). An interaction-based approach to enhancing secondary school instruction and student achievement. *Science*, *333*, 1034–1037.

Atteberry, A., & Bryk, A. S. (2011). Analyzing teacher participation in literacy coaching activities. *The Elementary School Journal*, *112*, 356–382.

Biancarosa, G., Bryk, A., & Dexter, E. R. (2010). Assessing the value-added effects of literacy collaborative professional development on student learning. *Elementary School Journal*, *3*(1), 7–34.

Blazar, D., & Kraft, M. A. (2015). Exploring mechanisms of effective teacher coaching: Results from two cohorts of an experimental evaluation. *Educational Evaluation and Policy Analysis*, *37*, 542–566.

Boaler, J., & Staples, M. (2008). Creating mathematical futures through an equitable teaching approach: The case of Railside School. *Teachers College Record*, *110*, 608–645.

Boston, M., & Smith, M. (2009). Transforming secondary mathematics teaching: Increasing the cognitive demands of instructional tasks used in teachers' classrooms. *Journal for Research in Mathematics Education*, *40*(2), 119–156.

Bryk, A. S., Gomez, L. M., Grunow, A., & LeMahieu, P. (2015). *Learning to improve*. Harvard Education Press.

Campbell, P. F., & Malkus, N. N. (2011). The impact of elementary mathematics coaches on student achievement. *The Elementary School Journal*, *111*(3), 430–454.

Cobb, P., Jackson, K., Henrick, E. C., & Smith, T. M., & MIST Team. (2018). *Systems for instructional improvement: Creating coherence from the classroom to the district office*. Harvard Education Press.

Coburn, C. E., & Penuel, W. R. (2016). Research–practice partnerships in education: Outcomes, dynamics, and open questions. *Educational Researcher*, *45*(1), 48–54.

Coburn, C. E., & Russell, J. L. (2008). District policy and teachers' social networks. *Educational Evaluation and Policy Analysis*, *30*(3), 203–235.

Coburn, C. E., Russell, J. L., Kaufman, J. H., & Stein, M. K. (2012). Supporting sustainability: Teachers' advice networks and ambitious instructional reform. *American Journal of Education*, *119*(1), 137–182.

Cohen, D. K., Raudenbush, S. W., & Ball, D. L. (2003). Resources, instruction, and research. *Educational Evaluation and Policy Analysis*, *25*, 119–142.

Darling-Hammond, L., Wei, R. C., Andree, A., Richardson, N., & Orphanos, S. (2009). *Professional learning in the learning profession* (p. 12). National Staff Development Council.

DeCaro, M. S., & Rittle-Johnson, B. (2012). Exploring mathematics problems prepares children to learn from instruction. *Journal of Experimental Child Psychology*, *113*(4), 552–568.

Desimone, L. M. (2009). Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. *Educational Researcher*, *38*(3), 181–199.

Desimone, L. M., & Garet, M. S. (2015). Best practices in teacher's professional development in the United States. *Psychology, Society, & Education*, *7*(3), 252–263.

Desimone, L. M., & Pak, K. (2016). Instructional coaching as high-quality professional development. *Theory into Practice*, *56*(1), 3–12.

Domina, T., Lewis, R., Agarwal, P., & Hanselman, P. (2015). Professional sense-makers: Instructional specialists in contemporary schooling. *Educational Researcher*, *44*(6), 359–364.

Duckworth, A. L., Tsukayama, E., & May, H. (2010). Establishing causality using longitudinal hierarchical linear modeling: An illustration predicting achievement from self-control. *Social Psychological and Personality Science*, *1*(4), 311–317.

Fennema, E., & Franke, M. L. (1992). Teachers' knowledge and its impact. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 147–164). Macmillan.

Fernandez, C., & Yoshida, M. (2004). *Lesson study: A Japanese approach to improving mathematics teaching and learning*. Erlbaum.

Foster, D., & Noyce, P. (2004). The mathematics assessment collaborative: Performance testing to improve instruction. *Phi Delta Kappan*, *85*(5), 367–374.

Franke, M. L., Turrou, A. C., Webb, N. M., Ing, M., Wong, J., Shin, N., & Fernandez, C. (2015). Student engagement with others' mathematical ideas: The role of teacher invitation and support moves. *The Elementary School Journal*, *116*(1), 126–148.

Gallucci, C., Van Lare, M. D., Yoon, I. H., & Boatright, B. (2010). Instructional coaching: Building theory about the role and organizational support for professional learning. *American Educational Research Journal*, *47*(4), 919–963.

Garet, M. S., Porter, A. C., Desimone, L., Birman, B. F., & Yoon, K. S. (2001). What makes professional development effective? Results from a national sample of teachers. *American Educational Research Journal*, *38*(4), 915–945.

Garet, M. S., Wayne, A., Stancavage, F., Taylor, J., Eaton, M., Walters, K., Song, M., Brown, S., Hurlburt, S., Zhu, P., Sepanik, S., Doolittle, F., & Warner, E. (2011). *Middle school mathematics professional development impact study: Findings after the second year of implementation* (NCEE 2011-4024). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

Gibbons, L. K., & Cobb, P. (2016). Content-focused coaching practices implicated in designing potentially productive coaching activities to support teachers' learning. *Elementary School Journal*, *117*(2), 237–260.

Gravani, M. (2007). Unveiling professional learning: Shifting from the delivery of courses to an understanding of the processes. *Teaching and Teacher Education*, *23*, 688–704.

Gravemeijer, K. (2004). Local instruction theories as means of support for teachers in reform mathematics education. *Mathematical Thinking and Learning*, *6*(2), 105–128.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, *8*(1), 23–34.

Hiebert, J. (2003). What research says about the NCTM Standards. In J. Kilpatrick, W. G. Martin, & D. Schifter (Eds.), *A research companion to principles and standards for school mathematics* (pp. 5–23). National Council of Teachers of Mathematics.

Hiebert, J., & Grouws, D. A. (2007). The effects of classroom mathematics teaching on students' learning. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 371–404). Information Age.

Hiebert, J., Morris, A. K., & Spitzer, S. M. (2018). Diagnosing learning goals: An often-overlooked teaching competency. In T. Leuders, K. Philipp, & J. Leuders (Eds.), *Diagnostic competence of mathematics teachers: Unpacking a complex construct in teacher education and teacher practice* (Vol. 11, pp. 193–206). Springer.

Hiebert, J., & Wearne, D. (1993). Instructional tasks, classroom discourse, and students' learning in second-grade arithmetic. *American Educational Research Journal*, *30*(2), 393–425.

Hill, H. C. (2009). Fixing teacher professional development. *Phi Delta Kappan*, *90*(7), 470–476.

Hopkins, M., Spillane, J. P., Jakopovic, P., & Heaton, R. M. (2013). Infrastructure redesign and instructional reform in mathematics: Formal structure and teacher leadership. *The Elementary School Journal*, *114*(2), 200–224.

Hopkins, M., & Woulfin, S. L. (2015). School system (re) design: Developing educational infrastructures to support school leadership and teaching practice. *Journal of Educational Change*, *16*(4), 371–377.

Jackson, K., Garrison, A., Wilson, J., Gibbons, L., & Shahan, E. (2013). Exploring relationships between setting up complex tasks and opportunities to learn in concluding whole-class discussions in middle-grades mathematics instruction. *Journal for Research in Mathematics Education*, *44*, 646–682.

Kapur, M. (2012). Productive failure in learning the concept of variance. *Instructional Science*, *40*(4), 651–672.

Kapur, M. (2014). Productive failure in learning math. *Cognitive Science*, *38*(5), 1008–1022.

Kennedy, M. M. (2016). How does professional development improve teaching? *Review of Educational Research*, *86*(4), 945–980.

Killion, J. (2008). *Assessing impact: Evaluating staff development*. Corwin Press.

Killion, J. (2012). Coaching in the K-12 context. In J. Fletcher & C. A. Mullen (Eds.), *The SAGE handbook of mentoring and coaching in education* (pp. 273–295). SAGE.

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.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*, Article 863.

Lampert, M. (2001). *Teaching problems and the problems of teaching*. Yale University Press.

Lewis, C. (2002). *Lesson study: A handbook of teacher-led instructional change*. Research for Better Schools.

Lewis, C., & Tsuchida, I. (1997). Planned educational change in Japan: The shift to student-centered elementary science. *Journal of Educational Policy*, *12*, 313–331.

Lewis, C., & Tsuchida, I. (1998). A lesson is like a swiftly flowing river: Research lessons and the improvement of Japanese education. *American Educator*, *22*(4), 12–17, 50–52.

Mangin, M. M., & Dunsmore, K. (2014). How the framing of instructional coaching as a lever for systemic or individual reform influences the enactment of coaching. *Educational Administration Quarterly*, *51*, 179–213.

Matsumura, L. C., Garnier, H. E., & Resnick, L. B. (2010). Implementing literacy coaching: The role of school social resources. *Educational Evaluation and Policy Analysis*, *32*(2), 249–272.

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.

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.

Matsumura, L. C., Sartoris, M., Bickel, D. D., & Garnier, H. E. (2009). Leadership for literacy coaching: The principal's role in launching a new coaching program. *Educational Administration Quarterly*, *45*(5), 655–693.

Neufeld, B., & Roper, D. (2003). *Coaching: A strategy for developing instructional capacity: Promises* *and practicalities*. Aspen Institute Program on Education.

Neuman, S. B., & Cunningham, L. (2009). The impact of professional development and coaching on early language and literacy instructional practices. *American Educational Research Journal*, *46*(2), 532–566.

Opfer, V. D., & Pedder, D. (2011). Conceptualizing teacher professional learning. *Review of Educational Research*, *81*(3), 376–407.

Otten, S., & Soria, V. M. (2014). Relationships between students' learning and their participation during enactment of middle school algebra tasks. *ZDM*, *46*(5), 815–827.

Penuel, W. R., Fishman, B. J., Cheng, B. H., & Sabelli, N. (2011). Organizing research and development at the intersection of learning, implementation, and design. *Educational Researcher*, *40*(7), 331–337.

Poglinco, S., Bach, A., Hovde, K., Rosenblum, S., Saunders, M., & Supovitz, J. (2003). *The heart of the matter: The coaching model in America*. CPRE Research Reports. http://repository.upenn.edu/cpre_researchreports/35

Polly, D. (2012). Supporting mathematics instruction with an expert coaching model. *Mathematics Teacher Education and Development*, *14*(1), 78–93.

Powell, D. R., Diamond, K. E., Burchinal, M. R., & Koehler, M. J. (2010). Effects of an early literacy professional development intervention on Head Start teachers and children. *Journal of Educational Psychology*, *102*, 299–312.

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

Rigby, J. G., Larbi-Cherif, A., Rosenquist, B. A., Sharpe, C. J., Cobb, P., & Smith, T. (2017). Administrator observation and feedback: Does it lead toward improvement in inquiry-oriented math instruction? *Educational Administration Quarterly*, *53*(3), 475–516.

Russell, J. L., Stein, M. K., Correnti, R., Bill, V., Booker, L., & Schwartz, N. (2017). Tennessee scales up improvement in math instruction through coaching. *The State Educational Standard*, *17*(2), 22–27.

Sailors, M., & Price, L. R. (2010). Professional development that supports the teaching of cognitive reading strategy instruction. *The Elementary School Journal*, *110*(3), 301–322.

Schwartz, D. L., & Martin, T. (2004). Inventing to prepare for future learning: The hidden efficiency of encouraging original student production in statistics instruction. *Cognition and Instruction*, *22*(2), 129–184.

Smith, M. S., Bill, V., & Hughes, E. K. (2008). Thinking through a lesson protocol: A key for successfully implementing high-level tasks. *Mathematics Teaching in the Middle School*, *14*(3), 132–138. (Reprinted in *Rich and engaging mathematical tasks: Grades 5-9*, by G. Lappan, M. S. Smith, & L. Jones, Eds., 2012, Reston, VA: National Council of Teachers of Mathematics.)

Stein, M. K., Correnti, R., Moore, D., Russell, J. L., & Kelly, K. (2017). Using theory and measurement to sharpen conceptualizations of mathematics teaching in the common core era. *AERA Open*, *3*(1). https://doi.org/10.1177/2332858416680566

Stein, M. K., Engle, R. A., Smith, M. S., & Hughes, E. K. (2008). Orchestrating productive mathematical discussions: Five practices for helping teachers move beyond show and tell. *Mathematical Thinking and Learning, an International Journal*, *10*(4), 313–340.

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

Stein, M. K., & Kaufman, J. H. (2010). Selecting and supporting the use of mathematics curricula at scale. *American Educational Research Journal*, *47*(3), 663–693.

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

Stein, M. K., & Meikle, E. (2017). The nature and role of goals in mathematics education. In D. Spangler & J. Wanko (Eds.), *Research companion to principles to actions* (pp. 1–11). National Council of Teachers of Mathematics.

Stigler, J. W., & Hiebert, J. (2004). Improving mathematics teaching. *Educational Leadership*, *61*(5), 12–17.

Tarr, J. E., Reys, R. E., Reys, B. J., Chavez, O., Shih, J., & Osterlind, S. J. (2008). The impact of middle grades mathematics curricula and the classroom learning environment on student achievement. *Journal for Research in Mathematics Education*, *39*, 247–280.

Webb, N. M., Franke, M. L., Ing, M., Wong, J., Fernandez, C. H., Shin, N., & Turrou, A. C. (2014). Engaging with others' mathematical ideas: Interrelationships among student participation, teachers' instructional practices, and learning. *International Journal of Educational Research*, *63*, 79–93.

Webster-Wright, A. (2009). Reframing professional development through understanding authentic professional learning. *Review of Educational Research*, *79*(2), 702–739.

Woulfin, S. L., & Rigby, J. G. (2017). Coaching for coherence: How instructional coaches lead change in the evaluation era. *Educational Researcher*, *46*(6), 323–328.

Authors

JENNIFER LIN RUSSELL, PhD, holds a joint appointment at the University of Pittsburgh as Professor and Chair of Educational Foundations, Organizations, and Policy in the School of Education and Senior Scientist and associate director at the Learning Research and Development Center. Her research examines policy and other educational improvement initiatives through an organizational perspective. Current work examines the way networks and other forms of research–practice partnerships are organized to accelerate systemic improvements that help educators address persistent problems of practice.

RICHARD CORRENTI, PhD, is an associate professor and research scientist at the University of Pittsburgh. His research interests center on measurement and determinants of teaching and how to improve teaching practice at-scale. His current projects focus on analyzing teaching development in the context of interventions such as literacy and mathematics coaching.

MARY KAY STEIN, PhD, holds a joint appointment at the University of Pittsburgh as professor of learning sciences and policy and senior scientist at the Learning Research and Development Center. Her research focuses on classroom-based mathematics teaching and the ways in which policy and organizational conditions shape teachers' practice. Her current work examines the role that cognitive mechanisms play in supporting teacher learning of a set of practices associated with productive discussions in mathematics and literacy classrooms.

ALLY THOMAS, PhD, is the senior director of Medicare STARs at UPMC Health Plan. Her work focuses on leading quality improvement across the health care system to improve the health quality, outcomes, and affordability for Medicare members.

VICTORIA BILL is a senior fellow at the Institute for Learning at the University of Pittsburgh. She is a designer and facilitator of research-based professional development with a specialization in mathematics coaching, curriculum, assessment and intervention as well as a background in school leadership and improvement science.

LAURIE SPERANZO is a math fellow at the Institute for Learning at the University of Pittsburgh. She focuses on creating and implementing professional development materials for math teachers, coaches, and administrators, particularly with a focus on equitable instructional practice in mathematics classrooms.

Manuscript received July 30, 2018
First revision received April 29, 2019
Second revision received October 19, 2019
Third revision received January 28, 2020
Fourth revision received May 21, 2020
Accepted June 8, 2020