ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 19-119

Developing Ambitious Mathematics Instruction Through Web-Based Coaching: A Randomized Field Trial

Matthew A. Kraft
Brown University

Heather C. Hill
Harvard University

This paper describes and evaluates a web-based coaching program designed to support teachers in implementing Common Core-aligned math instruction. Web-based coaching programs can be operated at relatively lower costs, are scalable, and make it more feasible to pair teachers with coaches who have expertise in their content area and grade level. Results from our randomized field trial document sizable and sustained effects on both teachers' ability to analyze instruction and on their instructional practice, as measured the Mathematical Quality of Instruction (MQI) instrument and student surveys. However, these improvements in instruction did not result in corresponding increases in math test scores as measured by state standardized tests or interim assessments. We discuss several possible explanations for this pattern of results.

VERSION: August 2019

Kraft, Matthew, and Heather Hill. (2019). Developing Ambitious Mathematics Instruction Through Web-Based Coaching: A Randomized Field Trial. (EdWorkingPaper: 19-119). Retrieved from Annenberg Institute at Brown University: http://www.edworkingpapers.com/ai19-119

Developing Ambitious Mathematics Instruction Through Web-Based Coaching: A Randomized Field Trial

Matthew A. Kraft
Brown University

Heather C. Hill
Harvard University

Updated: August 2019

Abstract

This paper describes and evaluates a web-based coaching program designed to support teachers in implementing Common Core-aligned math instruction. Web-based coaching programs can be operated at relatively lower costs, are scalable, and make it more feasible to pair teachers with coaches who have expertise in their content area and grade level. Results from our randomized field trial document sizable and sustained effects on both teachers' ability to analyze instruction and on their instructional practice, as measured the Mathematical Quality of Instruction (MQI) instrument and student surveys. However, these improvements in instruction did not result in corresponding increases in math test scores as measured by state standardized tests or interim assessments. We discuss several possible explanations for this pattern of results.

Matthew A. Kraft (corresponding author), Department of Education, P.O. Box 1938, Brown University, Providence, RI 02912. Email: mkraft@brown.edu; Heather C. Hill, 445 Gutman Library, Harvard Graduate School of Education, 6 Appian Way, Cambridge, MA 02138. Phone: 617-495-1898. Email: heather_hill@harvard.edu. This research was generously support by National Science Foundation (DRL-1348144). We are grateful to the many people who made this study possible, including Fallon Blossom, Samantha Booth, Mark Chin, Claire Gogolen, Corinne Herlihy, Dan McGinn, the MQI coaches, and numerous district personnel who generously donated their time. All errors and omissions are our own.

Developing Ambitious Mathematics Instruction through Web-Based Coaching:

A Randomized Field Trial

Collectively, public school districts invest tens of billions of dollars annually to improve classroom instruction, typically through teacher in-service training and professional development (Killeen, Monk, & Plecki, 2002; Miles, Odden, Fermanich, & Archibald, 2004; Jacob & McGovern, 2015). However, recent studies find mixed evidence regarding the impacts of professional development programs on instruction and student achievement. For instance, while some studies of STEM professional development programs find positive effects on student outcomes (Kisa, 2014; Roth et al., 2015; Penuel, Gallagher, & Moorthy, 2011; Roschelle et al., 2010), others find null or mixed results (Argentin, Pennisi, Vidoni, Abbiati, & Caputo, 2014; Jacob, Hill, Corey, 2017; Dominguez, Nicholls, & Storandt, 2006; Garet et al., 2011; Garet et al., 2016; Santagata et al., 2011). These results have caused some to question the value of investments in professional development as traditionally conceived (Jacob & McGovern, 2015).

Evidence to date suggests that teacher coaching programs may be an exception to these mixed and discouraging results. After small-scale experimentation in the 1980s and early 1990s, many scholars and practitioners advocated instructional coaching as a potentially successful workforce development strategy, leading to the growth of coaching programs in urban districts by the early 2000s (Neufeld & Roper, 2002; Russo, 2004). Results so far have been promising. A recent meta-analysis of 60 studies that used randomized control trials or rigorous quasi-experimental methods to evaluate teacher coaching programs found that, on average, the programs improved instructional quality by half a standard deviation and student achievement by almost one fifth of a standard deviation (Kraft, Blazar, & Hogan, 2018). However, this evidentiary base is largely limited to programs focused on literacy and teachers' general

pedagogical practice. In fact, there exists only one rigorous evaluation of math-specific coaching (Campbell & Malkus, 2011), despite the fact that over 18% of all public schools in the United States employ a math coach.[1]

To expand this evidence base, we describe and evaluate a web-based coaching program, MQI Coaching, that we designed to support teachers as they implement Common Core-aligned mathematics instruction. Our evaluation of MQI Coaching has several features that make it distinct from most prior studies of coaching programs. First, we provide a detailed theory of action based on evidence from the adult learning literature; in particular, we focus on calibrating teachers' views of instruction to our own, allowing their self-reflection to be more accurate and thus more powerful in driving change. Second, we collected an unusually rich set of data that enables us to test our theory of action, including evidence from teachers and coaches about the content of coaching sessions, evidence from students about their mathematics lessons, observations of instruction, and student test scores. Beyond presenting evaluation results, the empirical evidence we bring to bear can inform the theory behind coaching and contribute to the design of future professional development programs.

In what follows, we review the literature on math coaching models and describe the theory of action behind how MQI Coaching was designed to affect teacher practice and, ultimately, student achievement. We then describe the sample, randomization design, and how we operationalized MQI Coaching in this study. We next present findings on implementation fidelity and results from our block-randomized control trial evaluation from both the year in which teachers participated in coaching and the follow-up year after coaching activities had ended. By collecting and analyzing data in the follow-up year, we are able to assess whether any effects were sustained (or had potentially increased) when teachers could use their experiences to

inform their planning and instruction for a full academic year. We conclude with a discussion of the implications of our findings for research, policy, and practice.

Prior Literature on Coaching

In recent years, teacher professional development has moved from one-time workshops to professional learning opportunities that are practice-based, content-focused, collaborative, and offer ongoing support throughout the school year (e.g., Borko, 2004; Garet, Porter, Desimone, Birman & Yoon, 2001; Loucks-Horsley, Stiles, Mundry, Love & Hewson, 2010; Putnam & Borko, 2000). One key component of that effort has been the widespread adoption of teacher coaching. Although coaches can play many roles in schools, for this discussion we define coaching as characterized by a 1:1 teacher-coach relationship focused on supporting teachers' instructional improvement. Under this definition, coaches typically engage in activities such as modeling classroom practices, observing teachers' instruction, facilitating critical self-reflection, and providing direct feedback.

Coaching satisfies many of the criteria for professional learning recommended by scholars, making it a promising avenue for instructional improvement. It is individualized, intensive, sustained, context-specific, and focused (Kraft, Blazar, & Hogan, 2018). At the same time, there exists considerable variation in the practice of coaching. Some models entail direct coaching, in which coaches model desired practices and give teachers explicit advice and feedback about how to improve instruction. Other models entail more reflective coaching, in which coaches prompt teachers to analyze their own instruction and subsequently plan for improvement. Coaching programs may also be more structured—e.g., offering coaches and teachers routines and tools for use in their conversations—or less structured, as when coaches and teachers collaboratively decide on their goals, their processes, and their timeline.

To date, the research literature on coaching has focused overwhelmingly on reading and literacy coaching models, in large part because federal funds from the 1999 Reading Excellence Act and 2002's No Child Left Behind helped expand growth in this sector. Recent research also reflects the prominence of literacy coaching. Of the 40 causal evaluations of content-specific coaching models identified in a recent meta-analysis, all but five focused on reading (Kraft, Blazar, & Hogan, 2018). After several decades of development, many literacy models appear to be highly effective. Analyzing the recent causal literature on literacy coaching programs specifically, the authors found pooled effects of 0.51 standard deviations on teachers' instructional practice and 0.19 standard deviations on students' reading achievement.

In contrast, there exists only a small body of literature on math-specific coaching, containing only a single randomized field trial, Campbell and Malkus's (2011) study of a whole-school math coaching model. This program provided leadership and instructional coach training to elementary school teachers whose administrators nominated them to become full-time, site-based math coaches. In addition to working with individual teachers, math coaches supported their schools through a variety of roles, including assisting individual students, coordinating testing, and developing math curricula and programming. The authors found increasing positive effects of the site-based, whole-school coaching model on student achievement across the three years coaches worked with schools.

A more general line of inquiry examines the practices that math coaches use when working with teachers to support their instructional improvement. In Gibbons and Cobb's (2016) case study of one coach, the authors identified relatively directive coach activities such as setting short- and long-term goals for teacher learning. Mudzimiri and her colleagues (2014) found more diversity among coaching approaches, including some that capitalized on teacher reflection and

others that were more directive. This and other studies further underscore the importance of establishing rapport with teachers and convincing them of the efficacy of unfamiliar teaching techniques (Bengo, 2016; Gibbons & Cobb, 2016; Mudzimiri et al., 2014).

This brief review highlights two issues within the mathematics coaching literature. First, there exists very little evidence regarding the efficacy of mathematics-specific coaching programs, and no evidence on the efficacy of remote coaching in this subject. Second, we know very little about the question of whether coaching relationships in mathematics should lean toward being more teacher-driven or coach-directed. While we do not test this latter question directly, we do test a program that mixes teacher-driven reflection and planning with coach-driven norming of teachers' analysis of instruction. We explain this model in more detail next.

The MQI Coaching Model Theory of Action

MQI Coaching was co-developed over a several-year period by researchers at [blinded] and [blinded]. The model uses a well-established observational instrument, the Mathematical Quality of Instruction (MQI) (see Hill, Kapitula, Umland, 2011; Learning Mathematics for Teaching, 2011; Kelcey, Hill, & Chin, 2019 for a description of instrument validity and score validation efforts). The MQI offers items that capture 16 key practices for mathematics instruction, including precision in the use of mathematical language, connections between representations and solutions, and student mathematical communication, reasoning, and explanations (see Appendix A). Each item comes with four score points that provide descriptions of good, better, and best implementations of the practice. In MQI Coaching, the instrument structures teachers' and coaches' reflections on, and conversations about, short videos of math instruction. A central element of our theory of action was that if teachers could learn to analyze instruction using the MQI's items and practice descriptors, they would a) use more of the

practices contained in those items, and b) strengthen the quality of these practices as implemented in their classrooms. For instance, knowing that the MQI considers "student mathematical explanations" important, and reading the score points associated with the item should lead teachers to encourage more frequent, lengthy, and sophisticated student explanations. With this in mind, we designed "MQI coaching cycles" that consisted of teachers recording video of their own mathematics teaching, discussing and scoring video with their coaches along a limited number of MQI items, and then planning for improvement on those specific items.

Specifically, for this project we redesigned the coaching program described in the references above using insights from both older and more recent literatures on adult behavior change. First, we drew upon adult learning theory (Knowles, Holton, & Swanson, 2012; Merriam, 2001), which argues that adult learners have an independent self-concept and thus best self-direct their own learning; prefer problem-centered, applied, and immediately impactful approaches; and are internally motivated. For instance, we designed a coaching cycle that allows teachers to self-direct their learning by choosing the broad dimension, specific practice, and code to focus on for each coaching cycle. The coaching conversation itself, described below, relies heavily on teacher self-reflection and analysis of their own instruction. At the end of each coaching conversation, teachers and coaches plan immediate action steps, selecting one or two ways in which teachers would "elevate" their practice in the week following the conversation.

Second, our coaching model combines teacher reflection with *calibration*. Many argue that the former is ideally suited to improving complex practices such as teaching. As Schön (1983) described, teaching is a highly complex, contingent and thus uncertain practice. These features of teaching reduce the likelihood of identifying discrete problems with rational solutions, and render reflection on action a more adaptable and sustainable improvement

pathway. However, many have noted that teacher reflection is neither natural nor uniformly practiced (Valli, 1997). In our own prior research, for instance, we observed impacts of MQI professional learning communities on teachers' capacity to analyze video from our library, but no impacts on teachers' reflections on their own practice (Beisiegel, Mitchell, Hill, 2018). We also observed, during prior work, that teachers' reflections on their own instruction were often uncalibrated with observers' perceptions and with external standards—often, teachers believed themselves to be engaging students in reasoning or discussion when, in fact, they were not.

We interpret this as a teacher-focused version of educational psychologists' observation that less skilled individuals often mis-estimate or over-estimate their skills (Kruger & Dunning, 1999; Lichtenstein & Fischhoff, 1977; in education, see also Bridwell-Mitchell & Fried, 2018), perhaps because they have fewer meta-cognitive strategies to help them judge their skill levels. However, feedback on the accuracy of self-assessments can substantially improve those assessments (Lichtenstein & Fischhoff, 1977). Thus, our coaching program focuses on calibrating teachers' judgments of instructional quality with our own MQI "lens" on instruction. Our coaching model sought to achieve this goal through having teachers view and rate clips from our video library ("stock clips"), and through guided self-reflection using language from MQI score points and evidence from their own videos.

Third, our coaching model draws on the notion of routines and accountability to structure coach–teacher conversations. Feldman and Pentland (2003) described routines as "repetitive, recognizable patterns of interdependent actions, carried out by multiple actors" (p. 95). In education, Coburn and Russell (2008) and Horn and Little (2010) provide evidence that the use of well-crafted routines can increase the depth and analytic power of teachers' conversations with one another or with coaches. Sherer and Spillane's (2011) case study of a K–8 Chicago

school undertaking organizational reform suggests that schoolwide routines can focus attention on instructional practice and, critically, create accountability for change. Our coaching model reflects these ideas, in that we hold coaches accountable for enacting a well-specified routine during their coaching conversations with teachers (described below) and embed this conversation in a wider set of routines for teachers to follow. We also set expectations, to the extent their schedules allow, that teachers will engage in a coaching cycle as frequently as every two weeks. Our goal is to increase the interpersonal accountability between teachers and coaches by ensuring that teachers know they must take action as the next meeting with the coach draws nearer.

Taken as a whole, we expected this program to a) result in the enactment of a specific set of coach-teacher routines during coaching conversations, b) facilitate teacher reflection on their practice and calibration with our project's "lens" on mathematics instruction, and c) produce changes in observed teacher instructional practice and ultimately student outcomes. Thus, in this paper we ask:

1) Did coaches and teachers implement the MQI Coaching routines as intended?

2) Did MQI Coaching lead teachers to self-reflect, to calibrate their views on mathematics instruction with the MQI "lens", and to take immediate action to address target areas they identified as in need of improvement?

3) Did the MQI Coaching program improve teachers' instruction and/or student achievement?

We next describe the methods we used to answer these questions.

Methods

Setting and Sample

Districts. We partnered with two public school districts in the same Midwestern state to evaluate the efficacy of the MQI Coaching model. One was a large, urban district serving almost 80,000 students across more than 150 schools, with the vast majority of students from low-income families (83%) and families of color (86%). The second, a smaller suburban district, serves more than 15,000 students across 36 schools; over 70% are white and 37% come from low-income families.

Teachers. We recruited 142 upper elementary and middle school teachers from 51 schools to participate in the study, with roughly equal representation from each of the two districts. To be eligible, teachers had to teach full time in Grades 3–8. We recruited both subject matter generalists (mostly elementary) and subject matter specialists (mostly middle school). Across both districts, 15 participating teachers worked in in-district charter schools. Eleven teachers in the larger district worked in English/Spanish bilingual education schools.

Table 1 provides information about the backgrounds, prior training, and professional practices of participating teachers. The majority were white (80%), female (82%), and certified via traditional full-time teacher preparation programs (84%). Most (64%) held a graduate degree—typically a master's degree in education—although only a fraction had taken three or more college-level math courses. Teacher experience varied widely across the sample: 17% had taught 0–4 years, 44% had taught 5–15 years, and 39% had taught 16 years or more. Teachers who volunteered to participate in the study were largely representative of the mostly white, female, and relatively experienced workforces in these districts. For example, the average level of experience across both districts, weighted to reflect the proportion of teachers from each district in our sample, is 12.97 years, just under the sample average of 13.72 years.

Coaches. We recruited 24 expert MQI coaches with backgrounds as long-time MQI

raters, experienced classroom teachers, and/or instructional coaches. Among these coaches, 21 had prior experience as K–12 math teachers, and 19 had prior experience coaching, mentoring, or advising K–12 teachers in any subject. Half of the coaches had worked as math curriculum developers or college-level math professors. One third reported specific experience coaching teachers in math.

Prior to the intervention year, coaches passed an MQI certification exam and subsequently gained substantial experience scoring video using the MQI. During the implementation year, coaches participated in an intensive 15-hour initial training and attended monthly professional development sessions. The training focused on enacting the MQI Coaching routines described below, including logistics (helping teachers upload video; tagging video for discussion), basic elements of the MQI Coaching program, and specific instructions regarding conversational routines to use with teachers. Coaches also rehearsed coaching conversations in pairs. We developed the monthly professional development sessions based on coaches' feedback about the challenges they experienced working with teachers. Project staff listened to recorded coach–teacher sessions in order to monitor coaches' fidelity of implementation and to identify topics for the monthly sessions. Project staff also provided direct feedback to coaches who struggled to implement the model with fidelity. In particular, coaches whose coach-teacher discussions did not stay grounded in the routines we describe below (e.g., coaches used the time to offer advice or praise rather than analyzing video using the MQI) were brought back on track.

Coaches' characteristics, education, and professional experiences differed from the participating teachers they worked with in several important ways. As shown in Table 2, the coaches were predominantly women, but were less racially and ethnically diverse than the teachers they worked with—all but one coach was white. On average, coaches were slightly

younger and had fewer years of K–12 classroom teaching experience than the teachers. Nevertheless, coaches outperformed teachers on the Mathematical Knowledge for Teaching (MKT) assessment, a measure of the common and specialized knowledge used in teaching (see Ball, Thames, & Phelps, 2008). Coaches scored a full 1.23 standard deviations higher than teachers on the MKT, suggesting they had substantially stronger content knowledge. Coaches also attended more prestigious undergraduate institutions and had substantially more formal education and specific training in math than participating teachers.

Randomized Field Trial Design

In the summer of 2014, we randomly assigned the 142 participating teachers to receive MQI Coaching or to a control condition. Randomization ensures that the observable and unobservable characteristics of our volunteer sample of teachers are balanced in expectation across conditions. To further facilitate covariate balance, we blocked based on school type within districts—elementary, K–8, middle, and charter schools. We found no statistically significant differences between treatment and control teachers on any of the characteristics we measured, suggesting that these groups did not differ on observable measures due to random chance in finite samples (Table 1).

We paired the 72 teachers assigned to receive coaching with a trained and certified MQI coach based on grade-level expertise, preferred meeting times, and level of experience. We attempted to have coaches specialize in a single district to maximize their understanding of the context in which teachers worked. All but three coaches worked with two to four teachers (two worked with a single teacher, and one worked with six).

We pre-registered our data collection and analysis plan with the Institute for Education Sciences' What Works Clearinghouse Randomized Control Trial Registry (ID #491). We

collected data on participating teachers and their students for two years to assess the impact of MQI Coaching during implementation (2014–2015) and in the follow-up year (2015–2016), when teachers had the potential to benefit from the full year of training. Two participating teachers left their districts after randomization but before the school year began in 2014; eight others left after the end of the implementation year. This resulted in a potential analytic sample of 140 teachers in the implementation year and 132 in the follow-up year.

MQI Coaching Intervention

As noted above, we developed a theory of action that held that if teachers could learn to analyze instruction using the items and score points on the MQI, they would use this new understanding to improve their instruction, both by adopting more of the practices named on the MQI, and by implementing them at a higher level of quality. To accomplish this goal, we developed the coaching program itself based on principles from the adult learning literature, as well as the literature on the role of routines and calibration in improving practice. Here we describe how we wove these elements into the MQI Coaching Intervention.

Treatment teachers began their participation in MQI Coaching with a two-day summer training institute. Project staff introduced the MQI observation instrument, the coaching routines, and the video-recording technology and procedures. Then, at the start of the school year, each teacher-and-coach pair had a one-on-one introductory conversation during which they discussed the teacher's existing practice, her long-term plans for the year, and, more concretely, plans for the first coaching cycle. One goal of this conversation was to help the coach understand the teacher's motivation for change; another was to allow the teacher to begin to self-direct her own learning by choosing an initial focal dimension and one to two MQI codes within that dimension to work on during the next meeting. All coach meetings took place over the Adobe Connect web

platform.

During this initial two-day training, project staff also began the process of calibrating teachers to the MQI's instructional "lens." The MQI was developed by individuals who relied upon the research base in mathematics education to help analyze video from actual U.S. classrooms, as described in (Learning Mathematics for Teaching, 2011). Thus the instrument contains four dimensions, each containing multiple items drawn from the mathematics education research base and from the types of activities observed in actual classrooms: (a) *Richness of the Mathematics*, which captures the presence of disciplinary practices such as mathematical generalizations and multiple solution methods as well as mathematical sense-making activities; (b) *Common Core-Aligned Student Practices*, which captures students' mathematical reasoning, explanations, and communication, as well as the cognitive demands of classroom tasks; (c) *Working with Students and Mathematics*, which captures teachers' use of student ideas and teachers' remediation of student misconceptions; and (d) *Teacher Errors,* which captures any mathematical errors the teacher introduces into the lesson (see Appendix A for a brief description of the items within each dimension).

Each item on the MQI has four performance descriptors, providing item-specific behaviors or activities that would indicate that the item is not present (0); touched on briefly or superficially (1); enacted with a combination of strong and weak features (2); or enacted with strong features (3). For instance, for 'students communicate,' (0) indicates students contribute scant one-word answers during the segment; (1) indicates that students provide one or two-word answers consistently throughout the segment; (2) indicates occasional more substantive student contributions, such as presenting solution methods or asking a mathematical question; (3) indicates consistent student contributions throughout the segment. Teachers' scores on these

dimensions or combinations of these dimensions have predicted students' academic achievement gains in several studies (Hill, Kapitula, & Umland, 2011; Kelcey, Hill, & Chin, 2019; Garet et al., 2016).

During the initial two-day training, MQI staff began the process of calibrating teachers to the instrument. This mainly involved viewing and scoring video clips, then revealing the "master scores" for those clips (scores generated by expert raters) and discussing why those master scores applied.

Next, teachers and coaches began enacting the MQI coaching cycle (Figure 1). This cycle begins when a teacher chooses an MQI item to work on, then films a lesson (Step 1). Teacher choice is key for both, although in practice most teachers started with *Common Core-Aligned Student Practices* and then moved to either *Working With Students* or *Richness*. After receiving the video, the coach viewed the lesson and extracted two clips to share back with the teacher (Step 2). The coach also chose a stock clip, intended primarily to enhance teachers' calibration with the instrument. However, coaches also selected clips to model good practice in the area that the teacher was working on. Occasionally, coaches would select a clip that had similar problems to one of the teacher's clips, in order to give the teacher a chance to score and discuss those problems in a lower-stakes setting. Teachers watched all three clips offline (Step 3), then the teacher and coach met to discuss the clips and collaboratively plan how to "elevate" future instruction on those items. Teachers next returned to their classrooms and implemented their agreed-upon action plans. Teachers were asked to implement their action plans within two weeks, making the coach-teacher discussion immediately impactful, and also increasing teacher accountability for their plans.

Within the conversations described in Step 4, coaches and teachers enact another set of

routines designed to calibrate teachers to the MQI standards and to encourage self-reflection at their own practice. They begin by reviewing and scoring stock clips on the teacher-chosen items, which helps teachers recognize and understand the instructional practices the MQI prioritizes, and calibrate to the MQI's judgments about the quality of those practices. For instance, stock clips may show non-examples of practice (e.g., "here, there was no student communication"), along with good, better, and best examples of these practices, as spelled out in each MQI item's score points. Then, coaches and teachers move to a discussion of how the teacher in the stock video could have elevated her MQI score. The coach then asks the teacher to reflect on her own clips, and the process of analysis and elevation repeats. Coaches encourage teachers to take the lead, directing their learning and solving their own problems of practice. At the end of the analysis of her own clips, the teacher sets goals for the next filming cycle—specific activities she will engage in with the aim of changing her practice and improving her MQI score.

Although coach-teacher discussions were not tightly scripted, we did ask coaches to follow the routines described above in Step 4, and to use a common set of prompts (e.g., "How did you score this clip for *student explanations*? Why?") when discussing each item and clip. Coaches' own training and expertise came into use in several ways during these conversations. First, coaches provided feedback on the accuracy of teachers' stock-clip scores and, more gently, when teachers scored their own instruction. Second, when teachers discussed their plans for elevating their practice, coaches provided guidance or challenges to teachers' lines of thinking, typically by asking questions, but also by making suggestions about pedagogical practices to try or action steps to consider. However, a key philosophy of the program is that teachers take the primary role in driving their own learning through self-reflection. Finally, coaches used their experience working with teachers to build a trusting relationship with teachers.

We note that no teachers in this study chose errors as a topic for coaching cycles, despite the fact that other studies suggest that errors occur in about 68% of lessons (Hill, Litke, & Lynch, in press). When coaches did see a teacher error, the MQI Coaching protocol required them to address it. Coaches often did so subtly: by choosing a stock video with a similar error, then prompting the teacher to notice that error; showing the teacher her own error, then asking whether the instruction would lead to any student misconceptions or misunderstandings; or addressing the error in the context of the "elevating" part of the discussion (e.g., "to make your explanation score even higher, you could be more clear about the difference between an expression and an equation.") Coaches and program staff consulted regularly over how to address teacher errors, and coaches became quite adept at correcting them while still maintaining trust and open conversation.

Teachers and coaches engaged over the remainder of the academic year in the biweekly five-step coaching cycle outlined in Figure 1. At the end of this implementation year, we collected letters that treatment teachers wrote to themselves about the main takeaways from participating in the program. We returned these letters to them at the start of the following school year. This was the only form of additional treatment or support we provided to treatment teachers in the follow-up year.

Measures for Assessing Coaching Routines and Teachers' Engagement with Coaching

We collected several sources of data to allow us to answer our first two research questions, about whether teachers and coaches implemented the MQI Coaching routines as intended, and whether MQI Coaching prompted teachers to self-reflect, calibrate, and take action to improve their practice. Here we describe the instruments used to collect this process data, and the subsequent measures we constructed from these data.

Teacher baseline survey. At the time of enrollment of teachers into the program, we collected information on teacher demographics and educational background for use as controls in our analysis. We also asked teachers to answer sets of Likert-type items we developed to capture their openness to feedback, challenges with student behavior, and use of reform practices. We estimated teachers' scores on these three scales using item response theory graded response models. Among our sample of teachers, the three scales had Cronbach alpha reliabilities of 0.73, 0.84, and 0.87, respectively.[2] Finally, we included a measure of MKT, customized to teachers' grade level. The MKT alpha reliability was between 0.72 and 0.76 depending on the form administered. All 142 teachers in study completed the background survey.

Post-conversation coach survey. We developed an online survey to collect data on coaches' perceptions of the length, activities, focus, and quality of the 610 coaching sessions. Questions focused on routines addressed the focus of the coaching sessions and whether the coach and teacher completed the specific elements of the coaching cycle. The survey also captured the date and duration of each coaching session and information on any scheduling or technical difficulties. We asked coaches to report on the extent of teacher calibration with the MQI during each coaching session. Finally, we asked coaches to respond to a series of Likert-type questions about the degree to which teachers engaged in critical self-reflection and shouldered the work of reflection and planning during the coaching conversation. The coach survey also asked whether teachers implemented the action steps identified in their previous coaching cycle, the specificity and quality of action steps identified by teachers, and the overall quality of the coaching cycle. We report the frequencies of these individual survey items below.

Teacher end-of-year survey. In both the intervention year and the follow-up year, teachers completed a two-part online end-of-year survey developed by the research team. To

accommodate lesson reflections (described below), we administered part 1 and part 2 of the survey about two weeks apart. To ascertain whether treatment and control teachers differed in their experiences during the MQI Coaching year, this survey included a range of Likert-type questions about their experiences with professional development and exposure to MQI Coaching. Treatment teachers also responded to a set of open-ended questions about how, if at all, they changed their instruction due to MQI Coaching, and about any barriers to adopting new instructional practices that they faced. We collected survey responses from 119 of the 140 study teachers who taught in the participating districts in the intervention year (85.0%) and from 100 of the 132 study teachers who taught in the participating districts in the follow-up year (75.8%).

To assess teachers' calibration with the MQI, as well as the extent to which MQI Coaching induced critical self-reflection and planning for improvement, the survey asked teachers to reflect on a recently taught lesson and then to offer a short response to a five-minute stock clip of mathematics instruction. We expected that treatment-group teachers would incorporate more MQI-specific wording and topics into both types of reflections, would be more critical when reflecting on their own lesson, and would plan more changes to their own future instruction. Specific survey questions included, for the stock video clips, prompts eliciting teachers' views on the mathematics of the clip, the teaching in the clip, and any other topics of significance. Questions about the recently taught lesson elicited teacher critiques and thoughts on what they would change. Two coders blind to treatment condition scored each response based on the measures described below; raters reached 80% agreement before beginning to code; they double-coded all responses, then reconciled discrepant responses to arrive at a final single score.

We constructed three measures using teacher responses to questions about stock clips and their own lesson clips: (a) *MQI Language*, the mean number of responses that used MQI

language or concepts in the analysis of the stock video; (b) *Critique*, the mean number of things teachers identified as going well and not going well across the two lesson reflections; and (c) *Change*, the mean of a four-category ordinal measure capturing the number of things teachers would change across the two reflections (from 0 items to 3 items or more). We standardized these measures based on the control group mean in each year.

Measures for Assessing Effects on Instruction and Achievement

To address our third research question, we use both original data collection and district administrative data to assess the effect of coaching on teachers' instruction and students' academic achievement.

Classroom instruction ratings. During the follow-up year, we collected up to five classroom videos per teacher and scored them using the MQI instrument. We randomly assigned two trained raters who were blind to treatment status to watch and score seven-and-a-half minute lesson segments on each of 16 items using the Low (1) to High (4) scale. We settled on 7.5-minute segments, as opposed to longer or shorter segments, because raters reported that longer segments were too cognitively burdensome and because shorter segments meant significantly more scoring time and cost. Every 7.5-minute segment in each lesson was scored, as were final segments that were more than a minute long. We created an overall score for each MQI dimension by first averaging item scores across all clips from a teacher, and then taking the mean of these averages within domains for each teacher. We standardized all four measures based on control group means. For Richness, Common Core-Aligned Student Practices, and Working With Students, higher scores indicate stronger instruction; for Errors, higher scores indicate that teachers made more errors and, therefore, indicate worse performance. We estimated the reliability of these scales as 0.67, 0.74, 0.75, and 0.56, respectively, from intra-class correlation

coefficients, which capture the proportion of variance across lessons that is between-teachers after scaling our lesson error variance by the 4.8 average lessons we coded per teacher.

Student survey. In both years of the study, participating teachers in both the treatment and control condition administered a student survey that we developed to capture students' perceptions of the classroom practices targeted by coaching.[3] For example, items asked (in lay language) whether teachers requested student explanations, pushed them to use mathematical vocabulary, used pictures and diagrams in instruction, or provided opportunities for students to work through challenging content. We used these responses to construct a single scale we called *Ambitious Instruction*, borrowing language from Cohen's (2011) description of disciplinarily rich, student-centered instruction. We did so after a principal component analysis suggested our 11 focal items loaded onto one primary factor.

We constructed scores for this measure using a graded response model and standardized these scores with the control group mean in each year. We estimated the alpha reliability of the teacher-level *Ambitious Instruction* measure to be 0.59. As this suggests, the lower reliability of this measure limited our statistical power to detect smaller effects. We collected student survey responses from 120 of 140 study teachers who taught in the participating districts in the intervention year (85.6%) and 102 of 132 study teachers who taught in the participating districts in the follow-up year (77.2%).

State achievement tests. To assess program impact on student achievement, we collected student performance data for both state standardized tests and district-administered interim tests in math. The study took place during a transitional time for testing in the state where the participating districts were located. In 2014–2015, the intervention year, the state administered for the first time a computer-based assessment developed by the Smarter Balance

Assessment Consortium (SBAC) in Grades 3–8. The SBAC tests comprise both multiple choice and constructed-response items aligned with the Common Core State Standards (CCSS).[4] The following year, the state abandoned the SBAC, contracting instead with the Data Recognition Corporation (DRC) to develop and administer a new suite of tests in Grades 3–8. The new computer-based exams included multiple-choice and technology-enhanced (e.g., click and drag) items, but no constructed-response items. The DRC tests were aligned with a new set of state standards that were adopted after mounting political opposition to the CCSS and CCSS-aligned tests led the state to abandon them. In practice, multiple district officials suggested that the new state standards, although different in name, were quite similar to the CCSS.

Interim math assessments. We complemented these state assessments with student performance on the Measures of Academic Progress (MAP), developed by the Northwest Evaluation Association—a computer-based adaptive test that assesses math skills for students in Grades 2–12. The test is untimed and employs several item formats, including multiple choice and "drag and drop." Both districts administered the MAP assessment in math throughout the intervention year. In the follow-up year, the smaller suburban district switched to the Star test, developed by Renaissance Learning. Like the MAP, the Star test is a computer-based, adaptive assessment of math skills for students in kindergarten through Grade 12. We standardized all math test score measures by grade and year using scores from the full population of students across both districts.

Analytic Approach

We estimate treatment effects on teacher and student outcomes using ordinary least squares (OLS) regression and multilevel models, as described in our pre-analysis plan. We begin by fitting the following OLS model for teacher-level outcomes, where Y represents a given

outcome for teacher *j*:

$$Y_j = \beta Treat_j + \gamma X_j + \pi_b + \varepsilon_j \qquad (1)$$

Here, coefficient β on the indicator for whether a teacher was randomly offered the opportunity to participate in the MQI Coaching program, *Treat*, is our parameter of interest. β captures the intent-to-treat (ITT) effect of offering teachers MQI Coaching. In all models we include fixed effects for randomization blocks, π_b. In our preferred models, we also include a vector of teacher characteristics, X_j, to correct for potential imbalances across treatment and control groups caused by chance sampling differences or attrition. In addition to controls for gender, age, race, certification pathway, and an indicator for holding a graduate degree of any type, we also control for whether teachers held a master's degree in education, the number of mathematics content and methods courses they took (undergraduate or graduate level), their scores on the MKT assessment, and scales from survey items designed to capture their openness to feedback, challenges with student behavior, and use of reform practices. We estimate robust standard errors across all models for teacher-level outcomes. Although teachers in our study are clustered within schools, a sizable fraction were the only participating teachers in their schools, making a multi-level modeling approach unfeasible.

For our student survey outcome, we modify Equation 1, as we are able to directly model the clustered nature of the data where multiple students are nested within teachers. Thus, for student *i* with teacher *j,* we fit the following multi-level model:

$$Ambitious\_Instruction_{ij} = \beta Treat_j + \gamma X_j + \pi_b + (\upsilon_j + \varepsilon_{ij}) \qquad (2)$$

Our coefficient of interest remains β, the ITT effect of MQI Coaching on students' perceptions of their teachers' ambitious instruction in math.[5] We also include random effects for teacher, υ_j, which are orthogonal to *Treat* by construction.

We analyze student achievement outcomes using an augmented version of Equation 2 that includes controls for prior academic achievement and student characteristics as follows:

$$A_{ij} = \alpha V_{i,t-1} + \beta Treat_j + \delta W_i + \gamma X_j + \pi_b + (\upsilon_j + \varepsilon_{ij}) \qquad (3)$$

Here, *A* represents student achievement on the summative state or formative MAP achievement test. In addition to our controls for teacher covariates, we also include prior measures of achievement in math and reading on both the state test and the MAP, represented by the vector *V*. Controls for student characteristics, *W*, include indicators for gender, race, free or reduced-price lunch eligibility, limited English proficiency, special education services, and grade level.[6]

Treatment–Control Contrast

The magnitude of the ITT estimates we find will reflect the intensity of the contrast between the experiences of the treatment and control groups. One of the reasons we selected the two participating districts was because they did not already have a robust coaching program for math teachers. In fact, teachers' survey responses reported in Table 3 do indeed show that there were stark differences in exposure to coaching across the treatment and control groups. Very few teachers in the control group reported engaging in any type of frequent or intensive professional development focused on math instruction. As illustrated in Figure 2, 92% of treatment teachers reported that they received any type of coaching about once a month or more, compared to 14% of the control group (p<.001). Control group teachers rarely, if ever, received feedback from an evaluator, mentor, or peer teacher, or attended workshops related to their math instruction. Instead, they reported engaging in less formal collaborative activities related to math instruction with their peers. That some teachers in the control group received coaching and professional development in math is not a threat to our research design. Rather, it reflects the baseline conditions that determine the degree to which randomly offering MQI Coaching to treatment

teachers changed their professional development opportunities in meaningful ways.

Attrition in the Follow-up Year

Consistent with evidence of high rates of teacher movement generally (e.g. Atteberry, Loeb, & Wyckoff, 2017), we saw significant attrition from our sample. Ten of the 142 teachers in the original randomization sample left their districts before the start of the follow-up year. Twenty-one teachers no longer taught math in their original district, including some who left the district entirely, some who taught other subjects, and some who left the classroom for administrative positions. A total of 28 teachers no longer taught math in a tested grade (Grades 3–8).

Within-district teacher turnover does not pose a problem to our analyses, as we tracked and observed teachers who transferred between schools. We also find no difference in rates of teacher within-district retention across treatment and control groups as shown in Table 4. However, we do find that being randomly assigned to participate in MQI Coaching increased the likelihood that teachers taught math again in the follow-up year by 10.6 percentage points and taught math in a tested grade by 17.7 percentage points.

The differential attrition resulting from these treatment effects creates a challenge for estimating unbiased treatment effects in the follow-up year. We address this challenge by bounding our estimates in the follow-up year using extreme assumptions about dynamic differential attrition following Lee (2009). The intuition of this approach is as follows: we first assume that the treatment effect induced treatment teachers with the very highest (lowest) outcomes to remain in the study. We then systematically remove these treatment teachers at the upper (lower) tail of the distribution and re-estimate treatment effects. Removing treatment teachers with the highest outcome values produces our lower-bound estimate; removing

treatment teachers with the lowest outcome values produces our upper bound-estimate.[7] In addition, we explore whether teachers with certain types of characteristics were more likely to exit the study than others.

Findings

Did coaches and teachers implement the MQI Coaching routines as intended?

We begin to answer this question by describing teacher participation in MQI Coaching activities generally, and then examine conversational routines more specifically. Teacher participation was high overall, but variable. Of the 72 treatment teachers, 68 attended at least one day of the two-day summer institute, with 61 attending both days. During the 2014–2015 school year, 63 of 72 treatment teachers participated in at least one coaching session, with an average of 9.7 cycles among them. The majority of active treatment teachers met frequently with their coaches: 36 participated in 10 or more cycles, 18 completed between five and nine, and nine met between one and four times (Figure 3).

The high dosage of coaching cycles achieved our goal of frequent interactions between teachers and coaches. Over 68% of the coaching cycles occurred within three weeks of the previous cycle. Data also suggest teachers and coaches dedicated substantial time to engaging with each other during their conversations. As shown in Figure 4, coach–teacher video conferences ranged between 20 and 100 minutes, with an average length of just over an hour. Coaches judged there to be sufficient time to complete each step of the MQI Coaching cycle in 95% of the sessions.

Coaches reported implementing the core steps of the coaching routine with consistently high fidelity. Coaches and teachers reviewed and discussed the selected stock clip from our video library 89% of the time. They reviewed the first and second video clips selected from

teachers' recorded lessons 98% and 91% of the time, respectively. Coaches reported that about half (45%) of coaching cycles focused on items from the Common Core-Aligned Student Practices domain, followed by Richness (28%) and Working with Students (27%). In line with the discussion above, coaching cycles never explicitly focused on Errors.

Did teachers self-reflect, calibrate with the MQI, and take immediate action?

As noted in our theory of action, we expected teachers to critically self-reflect on their own practice during the enactment of these routines, to calibrate with the MQI's vision of instruction, and to plan for immediate improvements in their instruction. We describe evidence for each in turn. Coaches reported that teachers were engaged in critically analyzing their own instruction in 87% of the coaching sessions. However, coaches were less likely than teachers to report that teachers took primary responsibility for shaping the action steps. Coaches reported that in 36% of the cycles teachers took primary responsibility, in 38% teachers and coaches contributed equally, and in 26% coaches took primary responsibility. However, 41% of teachers reported that they took the primary responsibility for identifying action steps during coaching, 46% reported contributing equally, and only 13% said the coach took the primary role.

Coaches also reported the extent to which they believed teachers were calibrated with the MQI. Specifically, in 84% of the coaching sessions, they reported that teachers appeared to understand "well" the MQI scoring criteria they worked on. Coaches also reported agreeing with teachers' analyses of their own video clips in 92% of the coaching sessions. Notably, coaches had access to master scores for the clips they used, suggesting that teachers had a strong understanding of the MQI.

Our own analysis of teachers' lesson reflections suggests they were more calibrated to the MQI when viewing stock video than control group teachers were. In Table 5, we see that

coaching increased the frequency with which teachers used MQI-related language to analyze stock clips by about 1.1 standard deviations in Years 1 and 2. This translates to an approximate doubling, from one to two, of the number of MQI-related statements teachers made per lesson. However, in neither year did we find measurable effects on teachers' critiques of their own performance on two recently taught lessons, or the number of changes they planned after reflecting on their own lessons.

Finally, coaches felt that evidence in the subsequent video recording suggested that teachers had fully implemented the action plan from the previous cycle 66% of the time, and partially implemented the plan another 25% of the time. Teachers' own survey responses affirmed these perceptions: 87% of teachers reported that they often or always implemented the action steps they discussed with their coaches.

Did the MQI Coaching program improve teachers' instruction?

We report primary estimated effects on teachers' instruction in Table 6 for the implementation year (Panel A) and the follow-up (Panel B) year. We present estimates from both baseline models without controls as well as models in which we control for a range of teacher and, when applicable, student characteristics. Comparing estimates across both models illustrates the robustness of our estimates.

As judged by students, MQI Coaching improved teachers' instructional practice in the implementation year. Our preferred models, which include controls, estimate an effect of 0.22 standard deviations on students' assessments of teachers' Ambitious Instruction. This is equivalent to moving a teacher at 50th percentile of Ambitious Instruction to the 59th percentile. Our estimate of effects on Ambitious Instruction in the follow-up year are smaller and no longer statistically significant (0.08 standard deviations from our model with controls).

We find large effects on teacher instruction in the follow-up year on three of the four MQI dimensions: Richness, Working with Students, and Common Core-Aligned Student Practices. Our preferred estimates when controlling for teacher characteristics are 0.73 standard deviations for Richness, 0.47 standard deviations for Working with Students, and 0.61 standard deviations for Common Core-Aligned Student Practices.[8] Even if we assume that the treatment induced the very highest-performing treatment teachers to remain in the study and provide video-recordings of their classrooms, we still find meaningful and marginally significant effects of MQI Coaching on instructional practice in the follow-up year. The lower-bound estimate of MQI Coaching on Richness is 0.37 standard deviations and Common Core Practices is 0.34 standard deviations, while our estimate for Working with Students, at 0.24 standard deviations, is no longer statistically significant.

To help facilitate a clearer understanding of the magnitude of these effects, we re-estimated treatment effects using our preferred model, with controls, in a dataset consisting of raw MQI scores from every individual lesson segment (n=6,415). We converted these ordinal raw scores into a binary measure, where we code scores of Mid (2) or High (3) as 1, and scores of Not Present (0) or Low (1) as a zero. Conditional on teacher characteristics, we estimate that MQI Coaching increased the probability a treatment teachers' segment would score a Mid or High for Richness by 9.6 percentage points (p=.001) relative to a control group mean of 26%, a 37% increase. Effects on Working with Students were a 7.0 percentage point increase (p=.049), which translates to a 15% increase relative to the control group mean of 46%. Effects on Common Core Practices were a 9.2 percentage point increase (p=.001), or a 35% increase relative to the control group mean of 26%. Together, these results suggest that MQI Coaching had a sustained impact on teachers' delivery of high-quality mathematical instruction in the year

after they received coaching.

Teachers' own perceptions were consistent with these data. In response to open-ended questions on the follow-up year survey, teachers reported that they continued to use more sophisticated questioning techniques, encourage classroom discussion, and emphasize precision in mathematical language. However, they also noted that several factors constrained their persistence with MQI Coaching instructional practices, including less time for reflection, less time for classroom discussions, curricula that were out of alignment with the MQI approach, competing school responsibilities, competing district mandates and instructional guidance, students with behavioral and/or other special needs, and, in some cases, principals or peers who did not agree with the MQI approach. Teachers also noted that the loss of coaching sessions themselves meant they were no longer actively working on their practice.

Did the MQI Coaching program improve students' achievement?

We present estimates of the effect of MQI Coaching on student outcomes in the intervention year and follow-up year in Table 7. Across all models, we find no evidence of impacts on student achievement in either year of the study. Even with our more precise conditional estimates, however, we cannot rule out the possibility of small to moderate effects. Our 95% confidence intervals for effects on state math tests include effects as large as 0.10 standard deviations. Confidence intervals for effects on interim math tests include effect up to 0.12 standard deviations in the implementation year and 0.19 standard deviations in the follow-up year. Thus, our limited statistical power prevents us from rejecting the null hypothesis that the true effect in the population is zero, and from rejecting that the program has modest but meaningful effects on test scores.

Moderation and Robustness Tests

We conducted a range of additional analyses to 1) test for moderation effects, 2) examine whether "spillovers" from treated to control teachers attenuated our estimates, and 3) check the robustness of our results to differential attrition. As described in our Online Appendix, we find no evidence that program effects differed by teacher characteristics or that spillover effects or differential attrition pose a significant threat to validity of our findings.

Discussion and Conclusion

MQI Coaching provides one model for web-based coaching programs designed to strengthen the quality of teachers' math instruction. In this project, we combined the use of an observation instrument with supportive and reflective coaching. Teachers used self-captured video and the instrument to analyze their own instruction, and reflect on how to "elevate" their instruction on specific MQI items. Regular web-based meetings with coaches likely fostered a degree of informal accountability, helping teachers to stay engaged in the continuous improvement process. Participating teachers who volunteered for the study and were randomized to receive coaching were overwhelmingly receptive to and engaged in the coaching process.

Our evaluation found that many aspects of our theory of action were supported by evidence. Coaches reported that teachers engaged in critical analysis of their own instruction, took or shared responsibility for making a plan of action, and were generally calibrated to the MQI. More objective data, in the form of teachers' analysis of two stock clips, also suggests participating teachers were more calibrated to the MQI than control-group teachers. Coaches also reported that immediate-action plans made at the end of the coaching discussion were either partly or wholly enacted by the time of the next coaching conversation. However, we saw mixed results for instruction and student outcomes with moderate to large effects on teachers' instructional practices, but no detectable effects on student achievement. Our findings highlight

both the promise and tensions inherent in coaching programs.

An unanticipated finding from our study was that MQI Coaching increased the probability that teachers would continue teaching mathematics, and teaching mathematics in tested grades. Because we did not pre-register this analysis, we consider this finding exploratory. However, if confirmed in future studies, it has substantial implications for schools and districts. Teacher attrition from STEM subjects is significant (Ingersoll & Perda, 2009), producing shortages for mathematics teacher in many labor markets (Sutcher, Darling-Hammond & Carver-Thomas, 2016). The ten-percentage point increase in retention in mathematics and almost 18 percentage point increase in retention in mathematics for tested grades would be meaningful, particularly for small districts in tight labor markets. Reasons for the lower teacher attrition rate for MQI Coaching teachers may include an increased sense of classroom success (Johnson & Birkeland, 2003), extra effort on the part of principals to retain coached teachers, or even lower teacher stress because of the availability of resources for improving instruction.

What might explain our pattern of results?

We posit that two main factors led to success in the area of improving mathematics instruction. First, the MQI instrument provides specific indicators to teachers regarding what high-quality practice looks like. The MQI names 16 practices to engage in (or avoid, in the context of the Errors dimension), directing teachers' attention during lesson planning and instruction itself to these key elements of teaching. The MQI also provides examples of good, better and best instantiations of these elements. This level of explicitness at both the item and instrument level, we believe, led teachers to more clearly see the steps they needed to take to improve their practice.

Second, we believe the presence of the coach kept teachers focused on mathematics

instruction, encouraged self-reflection, and helped brainstorm ways to 'elevate' instruction. In particular, the use of talk routines during coaching conversations likely marshalled coach and teacher cognitive resources toward analyzing and improving instruction, rather than allowing the pair to fall into discussions aimed at strengthening their bond (e.g., chit-chat) or the coach to focus on demonstrating her expertise (e.g., when coaches deliver too-generous-amounts of advice). Similar successful routines also appear in other video-based coaching programs, such as My Teaching Partner (Allen et al., 2015). Future work should compare this form of highly structured coaching with less-structured coaching (e.g., programs in which the coach and teacher jointly determine structure and activity) to help guide future coaching initiatives.

We view the instructional changes caused by the MQI Coaching program as important outcomes in their own right. Coaching resulted in higher-quality instruction where students were given more opportunities to reason mathematically and make sense the mathematics. Transforming mathematics classrooms toward places where students think and reason has been a major reform goal for well over two decades, in part because discourse in such classrooms more closely resembles discourse in the discipline of mathematics, and because student thinking and reasoning is thought to prepare students to be more effective problem-solvers and critical thinkers as adults.

At the same time, these changes in teachers' instruction did not produce measurable improvements in student achievement on formative or summative math tests. There are several possible explanations for this pattern of results. It is possible MQI Coaching—and by extension, the practices it instilled in teachers' instruction—simply did not improve students' math skills. This is a serious issue for the mathematics education community; a major premise of scholars' work in this field is that higher quality mathematics teaching will lead to more student learning,

and neither this article nor similar recent studies (Garet et al., 2008; Garet et al., 2016) have been able to confirm that hypothesis. Because there are other program evaluations that provide more hope (e.g., Campbell & Malkus, 2011; Carpenter et al., 1989), a major task for scholars is to unpack when and how changes in instructional practices can lead to better outcomes for students.

It is also possible that improved math instruction strengthened students' abilities in ways not captured by the state standardized test or the interim assessments. Prior work (Kelcey, Hill, & Chin, 2019; Lynch et al., 2016) found that the relationship between instructional quality as measured by the MQI and student achievement varied by district—and specifically, the assessment used by the districts. Assessments that required students to engage in more cognitively demanding problem solving and explanations saw stronger relationships with MQI scores. This is in line with recent arguments that standardized achievement tests may not measure the thinking skills promoted by STEM projects, such as students' ability to conduct scientific investigations or make mathematical arguments (Sussman & Wilson, 2018).

Finally, it is possible that the effects on math achievement that resulted from MQI Coaching were too small to detect, given the limited power of our research design. Given the confidence intervals around our estimates (roughly -.10 to +.10), we can rule out medium to large effects. However, this study was not powered to detect effects smaller than 0.10 standard deviations. It is difficult to say with any certainty which explanation—power, the nature of standardized assessments, or lack of program efficacy—is most likely.

Program Costs

We provide information on program costs to allow policymakers to weigh these costs against the benefits of the program. We estimate that it cost approximately $4,000 per teacher to deliver MQI Coaching as part of this study. When we remove development- and research-related

costs, the estimate is closer to $3,500. These estimates are driven by three primary inputs: (a) coach compensation ($1,500 per teacher), (b) technology costs ($1,200 per teacher), and (c) costs for certifying, training, and supporting coaches ($500 per teacher). These costs are at the lower end of the range of prior estimates for site-based coaching models, with a substantially higher average number of coaching cycles per teacher relative to costs (Knight, 2012). We expect that on a per-cycle basis, web-based programs like MQI Coaching are likely to be more cost effective than site-based programs, even accounting for their additional technology requirements. Taking the retention effects at face value would also suggest the program prevented districts from having to hire about 7 new math teachers, with estimates for filling these vacant positions ranging between $10,000 to $20,000 dollars per teacher (Synar & Maiden, 2012; Watlington, Schockley, Guglielmino, & Felsher, 2010). This suggest that savings from increased teacher retention ($70,000 to $140,000) could reduce the net non-research-based cost of MQI Coaching ($233,000) by between 30% to 60%.

Looking forward

Developing, refining, and scaling coaching models takes time. This model relies on heavy structure through routines and an observation tool, and on teacher reflection. However, other math coaching models—including those with less formal structure and/or more coach direction of teacher practice—might achieve similar results in even more efficient ways. Compared to the decades-long history of literacy coaching and its rich evidentiary base, math coaching practice and research is still in its infancy. This study suggests that experimenting with new math coaching models and continuously refining existing models such as MQI Coaching is a worthwhile investment.

Endnotes

[1]Authors' calculations based on 2015–2016 National Teacher and Principals Survey data.

[2]We estimate this group-level reliability as the ratio of true variance over total variance $\frac{\text{Var}(\theta)}{\text{Var}(\hat{\theta})}$. To estimate the variance of the true theta scores $\text{Var}(\theta)$, we subtract the mean of the squared conditional standard errors of measurement (CSEMs) from the variance of the observed theta scores $\text{Var}(\hat{\theta})$.

[3]To minimize the risk of teachers influencing student answers, we made the surveys anonymous and provided each student with a sealable blank manila envelope in which to place the completed survey. We attempted to collect a systematic and representative sample of responses by instructing teachers to administer the surveys in the first and second math classes they taught, as well as by providing them with several weeks to administer the survey. This allowed them to obtain responses from students who were absent on the day the survey was administered. Teachers in Grades 3 and 4 were instructed to read the survey items and response anchors out loud for students. The full survey protocol and instrument are available upon request.

[4]The SBAC test administration in the state we studied did not utilize the adaptive nature of the online SBAC test or include any open-ended performance task items.

[5]The anonymous nature of the student survey precludes inclusion of student-level covariates.

[6]Valid prior measures of achievement are available for 72%–77% of students in the analytic samples for state standardized tests and 90%–93% for MAP tests. We estimate Equation 3 using multiple imputation, following Rubin (1987), in order to maintain a consistent sample across model specifications. We constructed 20 distinct data sets where missing data were imputed using student demographic characteristics and indicators for school assignment. Estimates represent the average effect across the 20 imputed data sets with their associated average standard errors corrected for the degrees of freedom used in the multiple imputation process.

[7]Lee (2009) bounds are particularly well suited for randomized trials with missing outcome data where no credible instruments exist and data are unlikely to be missing at random, conditional on a set of covariates. The Lee bounding approach assumes (a) that the predictor of interest is independent from the errors in the conventional outcome and selection models, and (b) monotonicity between treatment status and sample selection. The first assumption is assured by random assignment of treatment status; the second is commonly invoked and plausible in this context.

[8]These effects on teacher instruction as judged by MQI scores are subject to potential bias if teachers in the treatment group who were trained on the MQI selected days to record their instruction when they were delivering lessons highly aligned with MQI practices. While we cannot completely rule out this type of gaming, substantial amounts of it seem unlikely given that treatment teachers had no incentive to do so and submitted videos directly to the research team rather than to their coaches.

References

Argentin, G., Pennisi, A., Vidoni, D., Abbiati, G., & Caputo, A. (2014). Trying to raise (low) math achievement and to promote (rigorous) policy evaluation in Italy: Evidence from a large-scale randomized trial. *Evaluation Review, 38*(2), 99–132.

Atteberry, A., Loeb, S., & Wyckoff, J. (2017). Teacher churning: Reassignment rates and implications for student achievement. *Educational Evaluation and Policy Analysis*, *39*(1), 3-30.

Allen, J. P., Hafen, C. A., Gregory, A. C., Mikami, A. Y., & Pianta, R. (2015). Enhancing secondary school instruction and student achievement: Replication and extension of the My Teaching Partner-Secondary intervention. *Journal of Research on Educational Effectiveness*, *8*(4), 475-489.

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

Beisiegel, M., Mitchell, R., & Hill, H. C. (2018). The Design of Video-Based Professional Development: An Exploratory Experiment Intended to Identify Effective Features. *Journal of Teacher Education*, *69*(1), 69-89.

Bengo, P. (2016). Secondary mathematics coaching: The components of effective mathematics coaching and implications. *Teaching and Teacher Education*, *60*, 88–96.

Borko, H. (2004). Professional development and teacher learning: Mapping the terrain. *Educational Researcher*, *33*, 3-15.

Bridwell-Mitchell, E. N., & Fried, S. A. (2018). Learning One's Place: Status Perceptions and Social Capital in Teacher Communities. *Educational Policy*, 0895904818802117.

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

Campbell, P. F., & Malkus, N. N. (2014). The mathematical knowledge and beliefs of elementary mathematics specialist-coaches. *ZDM*, *46*(2), 213-225.

Coburn, C., & Russell, J. (2008). Getting the most out of professional learning communities and coaching: Promoting interactions that support instructional improvement. *Learning Policy Brief*, *1*(3), 1–5.

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

Dominguez, P. S., Nicholls, C., & Storandt, B. (2006). *Experimental methods and results in a study of PBS TeacherLine math courses*. Syracuse, NY: Hezel Associates.

Feldman, M. S., & Pentland, B. T. (2003). Reconceptualizing organizational routines as a source of flexibility and change. *Administrative Science Quarterly*, *48*(1), 94–118.

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., Heppen, J. B., Walters, K., Parkinson, J., Smith, T. M., Song, M., ... & Borman, G. D. (2016). *Focusing on mathematical knowledge: The impact of content-intensive teacher professional development* (NCEE 2016-4010). Washington, DC: National Center for Education Evaluation and Regional Assistance.

Garet, M. S., Wayne, A. J., Stancavage, F., Taylor, J., Eaton, M., Walters, K., ... & Sepanik, S. (2011). *Middle school mathematics professional development impact study: Findings after the second year of implementation.* Washington, DC: National Center for Education Evaluation and Regional Assistance.

Gibbons, L. K., & Cobb, P. (2016). Content-focused coaching: Five key practices. *The Elementary School Journal*, *117*(2), 237–260.

Hill, H.C., Kapitula, L.R. & Umland, K. L (2011). A validity argument approach to evaluating value-added scores. *American Educational Research Journal 48(3),* 794-831.

Hill, H. C., Litke, E., & Lynch, K. (in press). Learning lessons from instruction: Descriptive results from an observational study of elementary classrooms. *Teachers' College Record*.

Horn, I. S., & Little, J. W. (2010). Attending to problems of practice: Routines and resources for professional learning in teachers' workplace interactions. *American Educational Research Journal*, *47*(1), 181–217.

Jacob, A., & McGovern, K. (2015). *The mirage: Confronting the hard truth about our quest for teacher development.* Washington, DC: The New Teacher Project (TNTP).

Jacob, R. T., Hill, H. C., & Corey, D. (2017). The impact of professional development on teachers' mathematical knowledge for teaching, instruction, and student achievement. *Journal of Research on Educational Effectiveness, 10*, 379–407.

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

Kelcey, B., Hill, H. C., & Chin, M. J. (2019). Teacher mathematical knowledge, instructional quality, and student outcomes: a multilevel quantile mediation analysis. *School Effectiveness and School Improvement*, 1-34.

Killeen, K. M., Monk, D. H., & Plecki, M. L. (2002). School district spending on professional development: Insights available from national data (1992–1998). *Journal of Education Finance*, *28*, 25–49.

Kisa, Z. (2014). *A quasi-experimental study of the effect of mathematics professional development on student achievement* (Doctoral dissertation). Retrieved from http://d-scholarship.pitt.edu/22789/1/ZahidKisa_EDT_PDF.pdf

Knight, D. S. (2012). Assessing the cost of instructional coaching. *Journal of Education Finance*, 52-80.

Knowles, M. S., Holton E. F., III, & Swanson, R. A. (2012). *The adult learner*. New York, NY: Routledge.

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.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, *77*(6), 1121–1134.

Learning Mathematics for Teaching. (2011). Measuring the mathematical quality of mathematics teaching. *Journal for Mathematics Teacher Education 14(1),* 25-47.

Lee, D. S. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *The Review of Economic Studies*, *76*(3), 1071-1102.

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance*, *20*(2), 159–183.

Loucks-Horsley, S., Stiles, E., Mundry, S., Love, N., & Hewson, P. (2010). Designing professional development for teachers of science and mathematics. (3rd ed.). Thousand Oaks, CA: Corwin.

Lynch, K., Chin, M., & Blazar, D. (2017). Relationships between observations of elementary mathematics instruction and student achievement: Exploring variability across districts. *American Journal of Education*, *123*(4), 615-646.

Merriam, S. B. (2001). Andragogy and self-directed learning: Pillars of adult learning theory. *New Directions for Adult and Continuing Education*, *2001*(89), 3–14.

Miles, K. H., Odden, A., Fermanich, M., & Archibald, S. (2004). Inside the black box of school district spending on professional development: Lessons from five urban districts. *Journal of Education Finance*, *30*(1), 1–26.

Mudzimiri, R., Burroughs, E. A., Luebeck, J., Sutton, J., & Yopp, D. (2014). A look inside mathematics coaching: Roles, content, and dynamics. *Education Policy Analysis Archives*, *22*(53), 1–28.

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

Neufeld, B., & Roper, D. (2002). *Off to a good start: Year I of collaborative coaching and learning in the Effective Practice Schools*. Cambridge, MA: Education Matters.

Penuel, W. R., Gallagher, L. P., & Moorthy, S. (2011). Preparing teachers to design sequences of instruction in earth systems science: A comparison of three professional development programs. *American Educational Research Journal, 48*(4), 996–1025.

Putnam, R. T., & Borko, H. (2000). What do new views of knowledge and thinking have to say about research on teacher learning? *Educational Researcher, 29*, 4-15.

Roschelle, J., Shechtman, N., Tatar, D., Hegedus, S., Hopkins, B., Empson, S., ... & Gallagher, L. P. (2010). Integration of technology, curriculum, and professional development for

advancing middle school mathematics: Three large-scale studies. *American Educational Research Journal, 47*(4), 833–878.

Roth, K., Wilson, C., Taylor, J., Hvidsten, C., Stennett, B., Wickler, N., ... & Bintz, J. (2015, March). *Testing the consensus model of effective PD: Analysis of practice and the PD research terrain*. Paper presented at the International Conference of the National Association of Science Teacher Researchers, Chicago, IL.

Rubin, D. B. (1987). Multiple Imputation for Nonresponse in Surveys (Wiley Series in Probability and Statistics).

Russo, A. (2004). School-based coaching. *Harvard Education Letter*, *20*(4), 1–4.

Santagata, R., Kersting, N., Givvin, K. B., & Stigler, J. W. (2010). Problem implementation as a lever for change: An experimental study of the effects of a professional development program on students' mathematics learning. *Journal of Research on Educational Effectiveness, 4*(1), 1–24.

Schön, D. (1983). *The reflective practitioner.* New York, NY: Harper & Collins.

Sherer, J. Z., & Spillane, J. P. (2011). Constancy and change in work practice in schools: The role of organizational routines. *Teachers College Record*, *113*(3), 611–657.

Sussman, J., & Wilson, M. R. (2019). The use and validity of standardized achievement tests for evaluating new curricular interventions in mathematics and science. *American Journal of Evaluation*, *40*(2), 190-213.

Sutcher, L., Darling-Hammond, L., & Carver-Thomas, D. (2016). A coming crisis in teaching? Teacher supply, demand, and shortages in the US. Washington, DC: Learning Policy Institute.

Synar, E., & Maiden, J. (2012). A comprehensive model for estimating the financial impact of

teacher turnover. *Journal of Education Finance*, 130-144.

Valli, L. (1997). Listening to other voices: A description of teacher reflection in the United. *Peabody Journal of Education*, *72*(1), 67–88.

Watlington, E., Shockley, R., Guglielmino, P., & Felsher, R. (2010). The high cost of leaving: An analysis of the cost of teacher turnover. *Journal of Education Finance*, 22-37.

Tables

Table 1. *Teacher Characteristics*

	Full sample	Large urban district	Small suburban district	Treatment	Control	*p*-value (treatment vs. control)
Elementary School Teacher	0.71	0.71	0.71	0.71	0.71	0.92
Middle School Teacher	0.29	0.29	0.29	0.29	0.29	0.92
Male	0.18	0.15	0.20	0.15	0.20	0.47
Age (years)	40.99	39.96	42.09	40.48	41.51	0.47
Black	0.09	0.01	0.17	0.06	0.13	0.13
Hispanic	0.09	0.04	0.14	0.08	0.10	0.76
White	0.80	0.95	0.65	0.83	0.77	0.34
Experience (years)	13.72	13.71	13.73	14.35	13.07	0.38
Alternative certification	0.16	0.05	0.28	0.19	0.13	0.26
Undergraduate degree from very competitive institution	0.32	0.33	0.32	0.35	0.30	0.55
Undergraduate degree in mathematics	0.13	0.14	0.12	0.11	0.14	0.52
Undergraduate degree in education	0.51	0.59	0.42	0.51	0.50	0.85
Any graduate degree	0.64	0.58	0.71	0.65	0.63	0.79
Master's degree in education	0.47	0.41	0.54	0.43	0.51	0.31
Three or more advanced math courses	0.20	0.21	0.20	0.19	0.22	0.62
Three or more math content courses	0.42	0.36	0.49	0.46	0.38	0.31
Three or more math methods courses	0.26	0.18	0.35	0.29	0.22	0.37
Mathematical Knowledge for Teaching (SD)	0.00	0.17	-0.18	0.09	-0.10	0.18
n	142	73	69	72	70	

Note. All values represent proportions unless other units are indicated. SD = standard deviation. *p*-values associated with a significance test of the difference between a given characteristic across treatment and control groups, conditional on randomization blocks, with robust standard errors.

Table 2. *Coach Characteristics Compared To Teacher Characteristics*

	Coaches	Teachers	Difference	*p*-value
Male	0.17	0.18	0.00	0.98
Age (years)	38.33	40.99	-2.65	0.22
Black	0.00	0.09	-0.09	0.12
Hispanic	0.04	0.09	-0.05	0.42
White	0.96	0.80	0.16	0.06
Experience (years)	4.88	13.72	-8.85	0.00
Alternative certification	0.04	0.16	-0.12	0.14
Undergraduate degree from very competitive institution	0.63	0.32	0.30	0.00
Undergraduate degree in mathematics	0.29	0.13	0.16	0.04
Undergraduate degree in education	0.33	0.51	-0.17	0.12
Any graduate degree	0.92	0.64	0.28	0.01
Master's degree in math	0.13	0.01	0.12	0.00
Master's degree in education	0.75	0.47	0.28	0.01
Three or more advanced math courses	0.71	0.20	0.51	0.00
Three or more math content courses	0.46	0.42	0.04	0.72
Three or more math methods courses	0.38	0.26	0.12	0.24
Mathematical Knowledge for Teaching (SD)	1.23	0.00	1.23	0.00
Prior experience coaching	0.79			
Prior experience as K–12 math teacher	0.88			
Prior experience as a math curriculum developer/professional developer/mentor	0.50			
Prior experience as math coach	0.33			
n	24	142		

Note. All values represent proportions unless other units are indicated. SD = standard deviation. *p*-values associated with a significance test of the difference between a given characteristic across coaches and teachers with robust standard errors. The proportion of MKT items correct are among a subset of five items that were common across coaches and teachers

Table 3. *Control Teachers' Experience with Professional Development in Implementation Year*

	% who responded (n = 57)				
	Never	Once this year	About once a semester	About once a month	More than once a month
Received instructional coaching in math from any source	35%	28%	23%	7%	7%
Attended workshops or trainings about math instruction	23%	39%	25%	9%	5%
Collaboratively planned or debriefed about math instruction with other teachers	7%	9%	16%	30%	39%
Received feedback on math instruction as part of a formal or informal evaluation process	40%	39%	14%	5%	2%
Received feedback on math instruction from mentor/peer teachers in the district	51%	12%	12%	12%	12%

Note: Rows do not always sum to 100% due to rounding.

Table 4. *Effects of MQI Coaching on Teacher Retention and Assignment in Follow-Up Year*

	Teach in district	Teach math	Teach math in grade with high-stakes math test
Treat	0.059	0.106+	0.177**
	(0.043)	(0.057)	(0.064)
Constant (control group mean)	0.900***	0.799***	0.713***
	(0.030)	(0.041)	(0.046)
n	142	142	142

Note. Cells report regression coefficients with associated robust standard errors reported in parentheses. All models include randomization block fixed effects. High stakes math tests given in Grades 3–8. +p<.10, *p<.05, **p<.01, ***p<.001.

Table 5. *Effects of MQI Coaching on Teacher Reflection*

	n(teachers)	Unconditional	Controls
		Panel A: Implementation year	
MQI language (stock clip response)	119	1.110***	1.104***
		(0.219)	(0.238)
Critique (own clip reflection)	118	0.138	0.099
		(0.189)	(0.225)
Change (own clip reflection)	119	-0.137	-0.168
		(0.184)	(0.215)
		Panel B: Follow-up year	
MQI language (stock clip response)	100	1.001***	1.133***
		(0.269)	(0.308)
Critique (own clip reflection)	100	0.369	0.243
		(0.278)	(0.260)
Change (own clip reflection)	100	-0.074	-0.054
		(0.210)	(0.241)

Note. Cells report regression coefficients from separate models with associated robust standard errors reported in parentheses. All models include randomization block fixed effects. Controls include teacher gender, age, race, certification pathway, graduate degree, whether held a master's degree specifically in education, number of advanced math courses, math content courses, and math methods courses scores on MKT assessment, and scales from survey items designed to capture openness to feedback, challenges with student behavior, and use of reform practices. All teacher reflection outcomes measured in control-group standard deviations. +p<.10, *p<.05, **p<.01, ***p<.001.

Table 6. *Effects of MQI Coaching on Teacher Instruction*

Outcomes	*n*(students)	*n*(teachers)	Unconditional	Controls	Unconditional (lower bound)	Unconditional (upper bound)
			Panel A: Implementation Year			
Ambitious Instruction	3,252	120	0.171*	0.220**		
			(0.072)	(0.075)		
			Panel B: Follow-up Year			
Ambitious instruction	2,591	102	0.012	0.082	-0.110	0.121
			(0.087)	(0.087)	(0.085)	(0.092)
MQI Richness		104	0.819***	0.732**	0.366+	1.132***
			(0.232)	(0.249)	(0.209)	(0.220)
MQI Working with students		104	0.649**	0.466+	0.237	1.061***
			(0.224)	(0.241)	(0.209)	(0.209)
MQI Errors		104	0.213	0.283	-0.271+	0.394+
			(0.210)	(0.214)	(0.162)	(0.221)
MQI Common core-aligned student practices		104	0.700***	0.612**	0.342+	0.956***
			(0.194)	(0.207)	(0.178)	(0.190)

Note. Cells report regression coefficients from separate models with associated robust standard errors reported in parentheses. Standard errors for ambitious instruction reported in parentheses are from models with random teacher effects and idiosyncratic student-level errors. All models include randomization block fixed effects. Controls for teacher outcomes and ambitious instruction include teacher gender, age, race, certification pathway, graduate degree, whether held a master's degree specifically in education, number of advanced math courses, math content courses, and math methods courses taken, scores on MKT assessment, and scales from survey items designed to capture openness to feedback, challenges with student behavior, and use of reform practices. Ambitious instruction is measured in control group standard deviations. State math test and MAP math test are measured in standard deviations based on the full student population across both participating districts. +p<.10, *p<.05, **p<.01, ***p<.001.

Table 7. *Effects of MQI Coaching on Student Test Scores*

Outcomes	*n*(students)	*n*(teachers)	Unconditional	Controls	Unconditional (lower bound)	Unconditional (upper bound)
			Panel A: Implementation Year			
State math test	4,673	132	-0.055	-0.018		
			(0.082)	(0.042)		
MAP math test	5,160	136	-0.029	0.018		
			(0.084)	(0.034)		
			Panel B: Follow-up Year			
State math test	4,349	114	-0.017	-0.003	-0.167*	0.139
			(0.086)	(0.048)	(0.083)	(0.086)
MAP/Star math test	4,501	121	0.027	0.074	-0.109	0.126
			(0.087)	(0.059)	(0.076)	(0.085)

Note. Cells report regression coefficients from separate models with associated standard errors reported in parentheses. Standard errors reported in parentheses are from models with random teacher effects and idiosyncratic student-level errors. Lower and upper bound estimates are based on bounding approach developed by Lee (2009). State math test and MAP/Star math test measured in standard deviations based on the full student population across both participating districts. Effects on state and MAP math tests from controlled models are estimated using multiple imputation with 20 replication datasets to account for missingness on prior state and MAP test scores. Controls for state and MAP math tests include prior measures of achievement in math and reading on both the state test and MAP as well as indicators for gender, race, free- or reduced-price lunch eligibility, limited English proficiency, special education services, and grade level. See Table 6 notes for model details. +p<.10, *p<.05, **p<.01, ***p<.001.

Figures

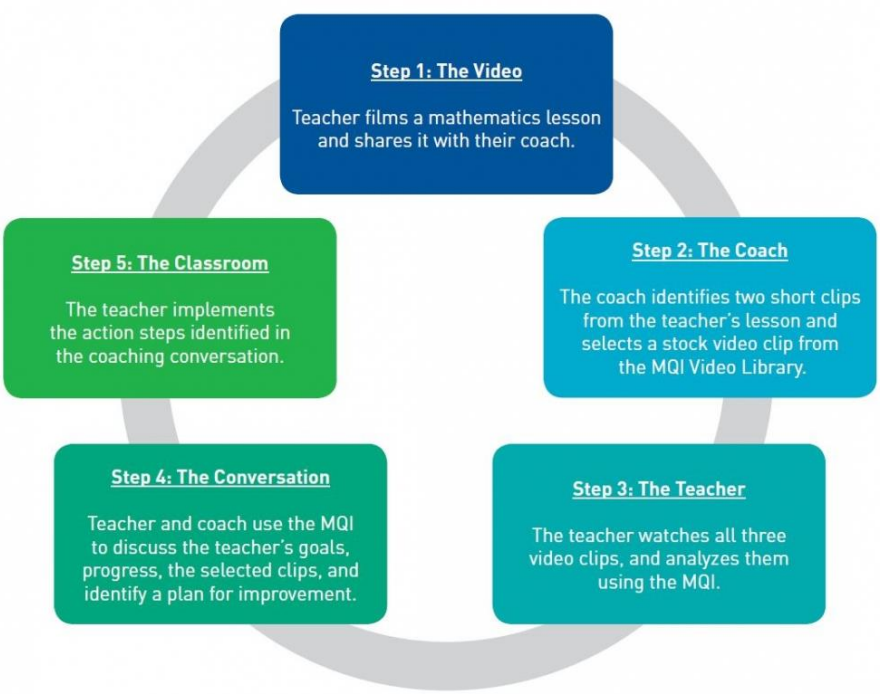

Figure 1. MQI Coaching cycle.

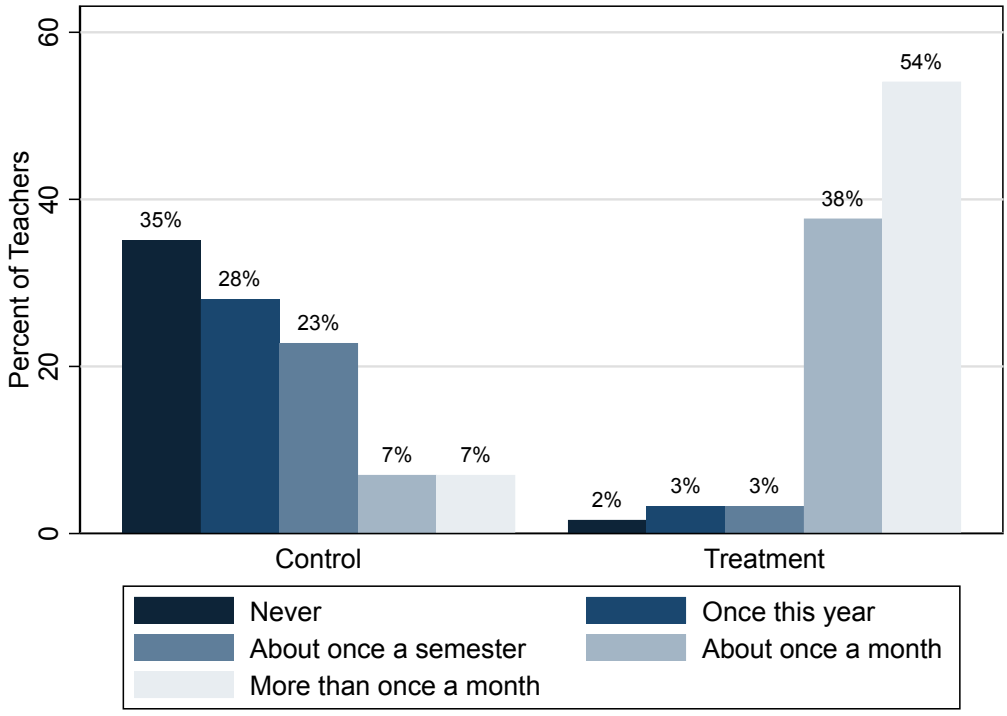

Figure 2. Treat-control contrast in the frequency teachers report engaging in instructional coaching in math.

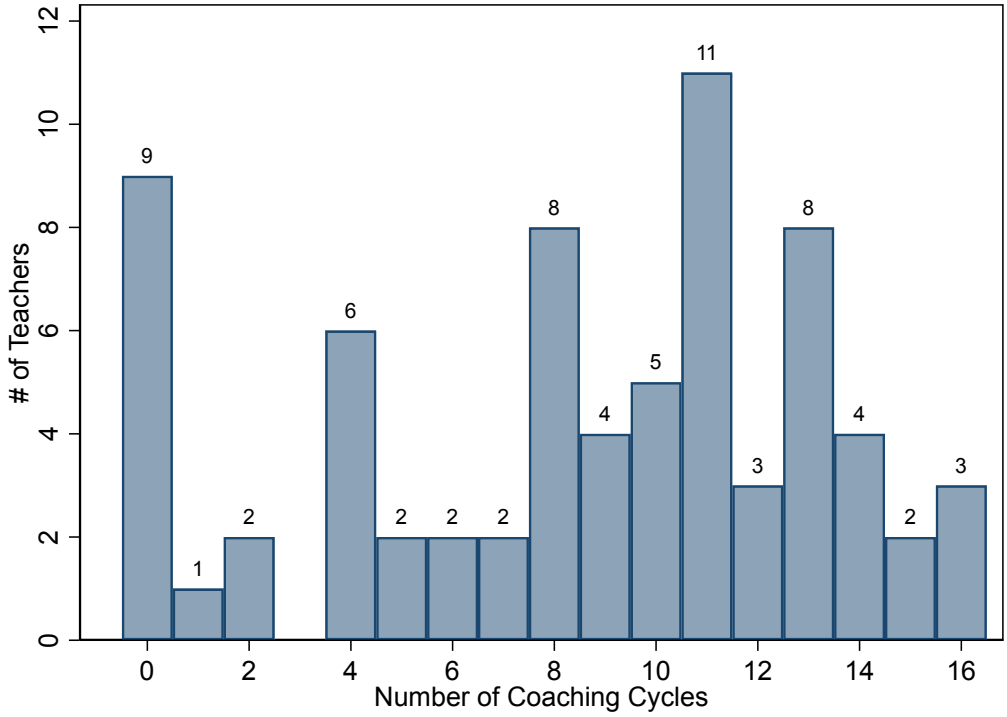

Figure 3. Number of coaching cycles completed across treatment teachers.

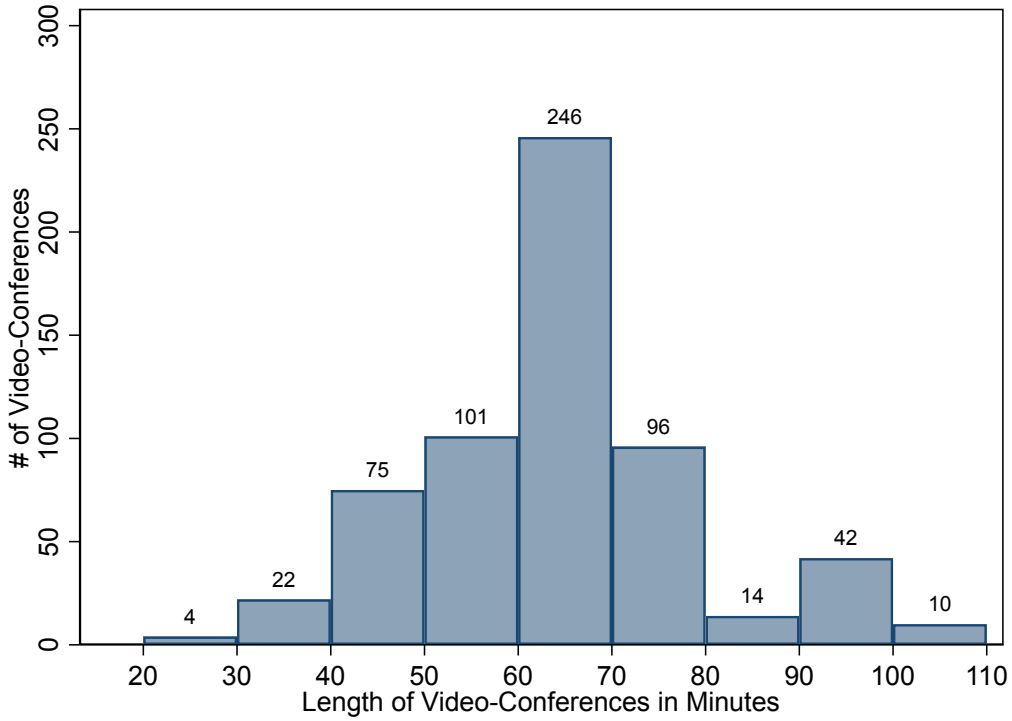

Figure 4. Length of coaching video-conferences between coaches and teachers.

Appendix Tables

Table A1. *Tests for Differential Attrition Across Treatment and Control Groups for Outcomes in Implementation Year and Follow-Up Year*

	Implementation year				Follow-up year				
	Video reflection measures	Ambitious Instruction	State math test	MAP math test	Video reflection measures	MQI rubric domains	Ambitious Instruction	State math test	MAP/ Star math test
Treat	-0.050	-0.064	-0.034	-0.004	-0.072	-0.156*	-0.157*	-0.177**	-0.106+
	(0.061)	(0.061)	(0.043)	(0.034)	(0.074)	(0.070)	(0.071)	(0.064)	(0.057)
Constant (control group mean)	0.187***	0.187***	0.088**	0.044+	0.332***	0.347***	0.361***	0.287***	0.201***
	(0.047)	(0.047)	(0.033)	(0.025)	(0.052)	(0.052)	(0.053)	(0.051)	(0.045)
n	142	142	142	142	142	142	142	142	142

Note. Robust standard errors reported in parentheses. All models include randomization block fixed effects.
+p<.10. *p<.05. **p<.01. ***p<.001.

Table A2. *Tests for Treatment Moderation on Teacher and Student Outcomes*

	MQI language (stock clip response)	Critique (own clip reflection)	Change (own clip reflection)	Richness	Working with students	Errors	CCASP	Ambitious instruction	State math test	MAP/ Star math test
					Panel A: Implementation year					
Treat*Small Suburban District	0.342	-0.587	0.261					-0.138	-0.008	0.064
	(0.450)	(0.380)	(0.374)					(0.135)	(0.068)	(0.064)
Treat*Experience	-0.001	-0.044+	0.025					-0.006	-0.005	-0.000
	(0.030)	(0.026)	(0.024)					(0.009)	(0.004)	(0.004)
Treat*MKT	-0.215	-0.254	-0.002					-0.058	0.010	0.057+
	(0.263)	(0.219)	(0.221)					(0.081)	(0.037)	(0.032)
Treat*Openness to Feedback	-0.156	-0.148	-0.299					0.117+	0.029	-0.009
	(0.206)	(0.206)	(0.209)					(0.069)	(0.037)	(0.030)
Treat*Challenges with Classroom Behavior	-0.169	-0.018	0.356+					-0.014	0.017	0.019
	(0.219)	(0.196)	(0.208)					(0.062)	(0.039)	(0.033)
Treat*Use of Reform Practices	-0.062	-0.274	-0.187					-0.039	-0.060	-0.106***
	(0.240)	(0.210)	(0.208)					(0.082)	(0.036)	(0.028)
					Panel B: Follow-up year					
Treat*Small Suburban District	-0.363	-0.509	-0.317	0.693	-0.225	0.022	0.366	0.087	-0.033	0.045
	(0.600)	(0.580)	(0.502)	(0.502)	(0.446)	(0.442)	(0.455)	(0.151)	(0.075)	(0.086)
Treat*Experience	0.018	-0.052	0.024	-0.004	0.020	0.004	-0.002	-0.002	-0.000	-0.002
	(0.041)	(0.038)	(0.029)	(0.033)	(0.032)	(0.025)	(0.026)	(0.010)	(0.005)	(0.006)
Treat*MKT	-0.160	-0.213	-0.074	0.423+	0.206	-0.015	0.065	0.062	0.020	-0.002
	(0.288)	(0.353)	(0.245)	(0.243)	(0.211)	(0.231)	(0.172)	(0.086)	(0.035)	(0.052)
Treat*Openness to Feedback	-0.209	-0.049	-0.004	-0.319	-0.202	-0.149	-0.078	0.033	0.044	0.033
	(0.229)	(0.292)	(0.264)	(0.245)	(0.221)	(0.182)	(0.211)	(0.077)	(0.036)	(0.042)
Treat*Challenges with Classroom Behavior	-0.022	0.114	-0.022	-0.069	-0.235	0.304	-0.054	-0.029	-0.018	0.010
	(0.227)	(0.249)	(0.237)	(0.251)	(0.246)	(0.223)	(0.247)	(0.082)	(0.040)	(0.042)
Treat*Use of Reform Practices	-0.619*	-0.049	-0.164	-0.325	-0.290	-0.246	-0.293	0.083	-0.002	-0.025
	(0.289)	(0.461)	(0.296)	(0.254)	(0.251)	(0.201)	(0.203)	(0.077)	(0.037)	(0.053)

Note. MKT = Mathematical Knowledge for Teaching. Cells report regression coefficients from separate models with associated robust standard errors reported in parentheses. Standard errors for student outcomes reported in parentheses are from models with random teacher effects and idiosyncratic student-level errors. Cells report estimates from separate models. All models include randomization block fixed effect and controls for teacher gender, age, race, certification pathway, graduate degree, whether hold a master's degree in education, number of advanced math courses, math content courses, and math methods courses, scores on MKT assessment, and scales from survey items designed to capture openness to feedback, challenges with student behavior, and use of reform practices. All moderators are measured in standard deviation units expect the indicator for small suburban district and experience which is measured in years. +p<.10, *p<.05, **p<.01, ***p<.001.

Table A3. *Tests of Differential Attrition from Study Across Teacher Characteristics for Follow-Up Year Outcomes*

	Teacher survey			MQI scores			Student survey			State test			Interim test		
	Stay	Exit	*p*-value	Stay	Exit	*p*-value	Stay	Exit	*p*-value	Stay	Exit	*p*-value	Stay	Exit	*p*-value
Elementary school teacher	0.77	0.57	0.66	0.76	0.58	0.82	0.76	0.57	0.99	0.73	0.64	0.73	0.74	0.52	0.70
Middle school teacher	0.23	0.43	0.66	0.24	0.42	0.82	0.24	0.43	0.99	0.27	0.36	0.73	0.26	0.48	0.70
Male	0.16	0.21	0.74	0.16	0.21	0.86	0.17	0.20	0.99	0.17	0.21	1.00	0.17	0.24	0.94
Age (years)	41.59	39.55	0.15	41.16	40.50	0.60	41.47	39.75	0.24	41.63	38.36	0.09	41.48	38.14	0.10
Black	0.08	0.12	0.73	0.08	0.13	0.60	0.08	0.13	0.62	0.08	0.14	0.30	0.08	0.14	0.44
Hispanic	0.07	0.14	0.33	0.08	0.13	0.56	0.08	0.13	0.60	0.10	0.07	0.56	0.09	0.10	0.85
White	0.83	0.74	0.45	0.83	0.74	0.57	0.82	0.75	0.62	0.81	0.79	0.79	0.81	0.76	0.77
Experience (years)	14.12	12.77	0.29	14.14	12.56	0.21	14.12	12.71	0.25	14.09	12.20	0.24	13.95	12.40	0.38
Alternative certification	0.17	0.14	0.17	0.16	0.16	0.26	0.18	0.13	0.07	0.18	0.11	0.12	0.17	0.10	0.05
Undergraduate degree from very competitive institution	0.33	0.31	0.65	0.35	0.26	0.21	0.34	0.28	0.26	0.33	0.29	0.65	0.33	0.29	0.61
Undergraduate degree in math	0.09	0.21	0.28	0.11	0.18	0.91	0.10	0.20	0.55	0.12	0.14	0.34	0.12	0.19	0.60
Undergraduate degree in educ.	0.49	0.55	0.22	0.50	0.53	0.39	0.50	0.52	0.43	0.54	0.39	0.25	0.52	0.43	0.62
Any graduate degree	0.66	0.60	0.31	0.66	0.58	0.20	0.66	0.60	0.36	0.65	0.61	0.73	0.65	0.57	0.46
Master's degree in education	0.49	0.43	0.69	0.49	0.42	0.67	0.49	0.43	0.69	0.49	0.39	0.65	0.49	0.38	0.78
# of advanced math courses	1.73	1.83	0.81	1.77	1.74	0.48	1.75	1.77	0.70	1.80	1.61	0.16	1.74	1.86	0.85
# of math content courses	2.42	2.40	0.91	2.43	2.37	0.64	2.45	2.33	0.37	2.42	2.39	0.74	2.42	2.38	0.69
# of math methods courses	2.19	2.24	0.65	2.21	2.18	0.86	2.19	2.25	0.59	2.23	2.11	0.40	2.21	2.19	0.84
Mathematical Knowledge for Teaching (SD)	-0.02	0.04	0.67	0.00	0.00	0.81	0.00	0.00	0.90	-0.05	0.20	0.68	-0.06	0.36	0.17
Openness to Feedback (SD)	-0.01	0.03	0.90	-0.01	0.03	0.98	-0.01	0.02	0.98	-0.04	0.16	0.16	-0.03	0.18	0.23
Challenges with Classroom Behavior (SD)	-0.07	0.18	0.23	-0.07	0.20	0.16	-0.07	0.19	0.17	-0.02	0.08	0.62	-0.02	0.11	0.58
Use of Reform Practices (SD)	0.06	-0.15	0.82	0.05	-0.13	0.96	0.02	-0.05	0.54	0.02	-0.09	0.72	0.02	-0.14	0.63
n	100	42		104	38		102	40		114	28		121	21	
p-value from joint F-test			0.76			0.75			0.40			0.57			0.51

Note. SD = standard deviation. *p*-values associated with a significance test of the difference between a given characteristic across teachers with valid outcome measures in the follow-up year and those who left the study, conditional on randomization block fixed effects, with robust standard errors. *p*-values from joint F-tests are from models using a binary indicator for missingness for a given outcome. Number of advanced, content, and math methods courses are measured on a scale from 1 (0 classes) to 5 (6+ classes).

Appendix A. Dimensions and Elements in the Mathematical Quality of Instruction (MQI) Instrument

Richness of the Mathematics

This dimension captures the depth of the mathematics offered to students. Rich mathematics focus either on the meaning of facts and procedures or on key mathematical practices. The dimension consists of the following elements:

- *Linking between representations*: Linking and connecting mathematical representations, ideas, and procedures.
- *Explanations*: Giving mathematical meaning to ideas, procedures, steps, or solution methods.
- *Mathematical sense-making*: Focusing on meaning in sustained ways
- *Multiple procedures or solution methods*: Considering multiple solution methods or procedures for a single problem.
- *Patterns and generalizations*: Using specific examples to develop generalizations of mathematical facts or procedures.
- *Mathematical language*: Using dense and precise language fluently and consistently during the lesson.

Common Core-Aligned Student Practices

This dimension captures evidence of students' involvement in cognitively activating classroom work. Attention here focuses on student participation in activities such as:

- *Students provide mathematical explanations* for an idea, procedure, or solution.
- *Student mathematical questioning and reasoning,* engage with important mathematical practices.
- *Students communicate about the mathematics of the segment* by asking mathematical questions, describing the meaning of a term, offering an explanation, discussing solution methods, commenting on the reasoning of others.
- *Task cognitive demand:* Students engage in task in which they think deeply and reason about mathematics.
- *Students work with contextualized problems*

Working with Students and Mathematics

This dimension captures evidence of teachers' use of students' misconceptions and mathematical ideas. Attention here focuses on two aspects of this work:

- *Remediation of student errors and difficulties*, where higher scores require teachers to conceptually address student misconceptions.
- *Teacher uses student contributions*, which captures the spectrum of ways students can participate in the class, from teachers who allow only one-word answers to teachers who weave student mathematical ideas at length into the development of the mathematics during the segment.

Teacher Errors

This dimension captures teacher errors or imprecision in language and notation, or the lack of clarity/precision in the teachers' presentation of the content. Attention here focuses on:

- *Mathematical content errors*, which records teachers' uncorrected errors with the content.
- *Imprecision in language and notation*, which records teachers' errors in notation, mathematical terms, and general language when used to describe math.
- *Lack of clarity*, which captures teachers' mathematics-related utterances that muddle, confuse, or distort the mathematical content.

Online Appendix

Moderation Analyses

We extend our primary analyses to include the exploratory moderation analyses outlined in our pre-analysis plan. Specifically, we examine whether treatment effects differ systematically by measures of teachers' experience, district, openness to feedback, challenges with student behavior, and use of reform practices described above. To do this, we adapt the relevant modeling approach (Equations 1–3) by adding the main effect of our moderator variable as well as an interaction term between *Treat* and the moderator of interest. The coefficient on this interaction term tests whether the effect of MQI Coaching differed across teachers based on their characteristics.

As shown in Appendix Table A2, our results suggest that effects of MQI Coaching were of similar magnitude across districts and for teachers with a wide range of prior background characteristics and teaching styles. Across outcomes, the coefficients associated with the interaction of *Treat* with a district indicator or teacher characteristics are of inconsistent signs and very rarely statistically significant. In fact, we find only two statistically significant estimates at the 0.05 level among the 94 interaction terms we tested. This is even fewer than we would expect due to Type I error alone. We interpret this as encouraging evidence that MQI Coaching may be effective at improving calibration and math instruction among teachers with a range of experience and pedagogical approaches.

Spillover

Our teacher-level randomization design maximized the statistical power of our analysis but created the possibility for within-school spillover effects across teachers assigned to treatment and control groups. In survey responses from 57 control group teachers at the end of

the intervention year, 75% reported knowing a teacher who received MQI Coaching and 19% reported ever talking with a teacher who received MQI Coaching. Six control group teachers (10.5%) reported collaboratively planning instruction with treatment group teachers, and five (8.8%) reported changing their math instruction based on ideas/techniques they learned from treatment teachers. Reports about exposure to MQI Coaching via treatment teachers in the follow-up year are quite similar. Overall, these findings suggest that spillover is not a major concern, but it might have attenuated our treatment estimates slightly.

Attrition

We further examine the potential threat posed by differential attrition across treatment and control groups in the follow-up year. We closely tracked reasons for attrition through exit surveys as well as informal communication with teachers and school administrators. The two most common reasons for exiting the study were that a teacher was no longer teaching math or had left their district entirely, as described above. Two other common reasons for attrition were a lack of time to participate (n=8; 6 treatment, 2 control) and a loss of interest in participating (n=6; 4 control, 2 treatment).

Differential attrition by itself does not mean that the characteristics of treatment and control groups are no longer equal in expectation. Although we cannot know if attrition was related to unobserved teacher characteristics correlated with outcomes, we can examine the relationships between our set of observed teacher characteristics and attrition. In Table A3, we report simple averages of 21 characteristics across teachers who are missing data for outcomes in the follow-up year and those who are not, as well as *p*-values from model-based *t*-tests of the group-mean differences after accounting for randomization blocks. We find no statistically significant differences across stayers and leavers on any measure across the five different follow-

up year outcomes. We fail to reject a null hypothesis of no relationship between all 21 measures and an indicator for exiting the study in joint significance tests across all five outcomes with *p*-values ranging from 0.51 to 0.76. These findings suggest that attrition from the study is driven by circumstances largely unrelated to teachers' observable characteristics, and thus it is unlikely to induce substantial bias in our follow-up year estimates.