Article

Journal of Teacher Education
2018, Vol. 69(1) 69–89

DOI: 10.1177/0022487117705096
journals.sagepub.com/home/jte
SAGE

The Design of Video-Based Professional Development: An Exploratory Experiment Intended to Identify Effective Features

Mary Beisiegel[1], Rebecca Mitchell[2], and Heather C. Hill[3]

Abstract

Although video cases and video clubs have become popular forms of teacher professional development, there have been few systematic investigations of designs for such programs. Programs may vary according to (a) whether teachers watch videos of their own/their peers' instruction, or whether teachers watch stock video of unknown teachers; and (b) whether discussions are led by trained facilitators or by participants themselves. Using a factorial design, we defined four treatment conditions based on these possibilities, then assigned three groups of teachers to each condition. Teachers watched, scored, and discussed mathematics instruction according to each treatment condition's protocol. Evidence from groups' conversations and teachers' video analyses and lesson reflections suggest that the teacher-led, own-video condition is slightly superior to the other conditions.

Keywords

mathematics teacher education, professional development, video analysis

Research into professional development has reached a turning point. A consensus achieved over the past decade about features that maximize the effectiveness of professional development (e.g., Borko, 2004; Desimone, 2009; Desimone & Garet, 2015; Kennedy, 1999) has been buffeted by recent and disappointing evidence from randomized clinical trials of well-designed, well-funded programs (e.g., Garet, Porter, Desimone, Birman, & Yoon, 2001; Garet et al., 2010; Santagata, Kersting, Givvin, & Stigler, 2011). Many of these programs incorporated the features thought critical to professional development success—a content focus, alignment with existing curriculum, length (multiple weeks or a school year) that allows teachers time to hone their craft, and use of artifacts of classroom practice such as videotape or student work—yet few have been found to affect teacher and student learning (Wayne, Yoon, Zhu, Cronen, & Garet, 2008). Elsewhere we have argued that researchers should complement such large-scale evaluations of specific programs with smaller scale experiments that systematically test program *features* (Hill, Beisiegel, & Jacob, 2013). By testing the efficacy of such features prior to scale-up, developers can both optimize their program design and contribute to the knowledge base about effective professional development.

Here we illustrate such an approach by describing an exploratory experiment conducted during the design phase of video-based mathematics teacher professional development. In this professional development program, teachers watched and analyzed video in pursuit of improved understanding of instruction. We argue that the research literature around such models invites two questions: What type of video is most effective? And what role should the facilitator play in the analysis and discussion of video? On the first count, teachers could analyze videos from one of two sources: videos of themselves and their peers teaching mathematics lessons, or stock video culled from contemporary video libraries. On the second count, facilitators could have a strong hand in governing discourse around the video, or could scaffold teachers to lead the discussion themselves. Though these are not the only questions in the field, we found them critical decision points in the design of our program, and expect that other professional developers may face similar uncertainties.

Below, we describe our research investigating the effects of video source and type of facilitation on professional development *processes*—the nature of teachers' conversations and analyses of their own videos—and proximal *outcomes*,

[1]Oregon State University, Corvallis, USA
[2]Pine Manor College, Chestnut Hill, MA, USA
[3]Harvard University, Cambridge, MA, USA

Corresponding Author:
Mary Beisiegel, Department of Mathematics, Oregon State University, Kidder Hall 368, Corvallis, OR 97331-4605, USA.
Email: mary.beisiegel@oregonstate.edu

including teachers' use of a framework introduced in professional development to structure their thinking about practice. In the sections that follow, we discuss the literature that sparked the questions outlined above, the design of our study and testing of each case, and the results of our inquiry.

Research on Mathematics Professional Development

A core component of our program is the use of classroom video in mathematics teacher professional development. This approach has increased over the last two decades, in part because advocates have noted that watching video clips of teaching provides teachers a common context for observing and analyzing instruction (Borko, Koellner, Jacobs, & Seago, 2010; Calandra, Brantley-Dias, Lee, & Fox, 2009; Hixon & So, 2009; Sherin, 2003) while preserving "the complexity and contextual nature of classroom teaching" (Kersting, 2008, p. 848). Studies have demonstrated that video use supports teacher noticing of salient aspects of practice, including student thinking and interactions between teachers and students (Borko, Jacobs, Eiteljorg, & Pittman, 2008; Mitchell & Ariemma-Marin, 2015; Sherin & van Es, 2005; Tripp & Rich, 2012; van Es & Sherin, 2010). Similarly, analysis of video appears to support teachers' learning about new instructional strategies (Sun & van Es, 2015), and teachers have reported applying in their lessons what they have learned through video analysis (Christ, Arya, & Chiu, 2014). As two examples, van Es and Sherin (2010) found that teachers who shared clips from their own lessons increasingly probed students' thinking and made students' ideas public, and Roth and colleagues (2011) found that teachers who analyzed their own lessons significantly increased their use of student thinking strategies.

Furthermore, researchers have found that teachers' ability to analyze video recordings of mathematics lessons is related to increases in teacher knowledge and students' learning of mathematical topics (Kersting, Givvin, Sotelo, & Stigler, 2010; Kersting, Givvin, Thompson, Santagata, & Stigler, 2012; Seago, Jacobs, Heck, Nelson, & Malzahn, 2014). As another example, in a rigorously designed trial, professional development in which science teachers used both content- and student-focused lenses to analyze video improved teachers' content and pedagogical content knowledge, changed classroom practices, and improved student achievement (Roth et al., 2011). However, whose videos are used and what type of facilitation is necessary for program success are still open questions in the research literature (Seidel, Stürmer, Blomberg, Kobarg, & Schwindt, 2011; Sherin & Han, 2004; Zhang, Koehler, & Lundeberg, 2015). Below, we describe research in these two areas.

Type of Video

One question concerns the source of video—that is, whether teachers should watch and analyze videos of their own teaching or videos culled from a library (Gaudin & Chaliès, 2015; Kleinknecht & Schneider, 2013). The decision to use teachers' own video carries several potential benefits and costs. Logically, one might expect that teachers analyzing their own teaching would increase the potential impact of the activity on practice, as they are critically viewing their own classrooms and receiving direct feedback regarding how to improve instruction (Harlin, 2014; Snoeyink, 2010). In addition, teachers viewing their own lessons find the experience more motivating and activating than teachers watching unknown teachers' lessons (Seidel et al., 2011). Moreover, if colleagues within schools or districts watch videos together, they can bring contextual information to bear on the instruction they analyze—for instance, how that instruction responds to the specific needs of students or to district instructional guidance (Borko et al., 2008; Goldsmith & Seago, 2011).

However, when viewing their own and their peers' lessons, teachers are often hesitant to analyze and make critical comments about each others' teaching (van Es, 2012; Zhang, Lundeberg, Koehler, & Eberhardt, 2011). Furthermore, teachers may "unconsciously develop a 'deceptive discourse'" to avoid conversations that lead to criticism of their own and others' teaching (Seidel et al., 2011, p. 261). Schools' "culture of nice," as some researchers have called it, may mean teachers engage in surface-level collaboration without gaining real insight into their teaching practices (MacDonald, 2011). Teachers may also experience problems with camera set-up, video capture, and processing (Zhang et al., 2011)—problems that may impede the successful use of the own-video condition at scale.

The use of stock video also has potential benefits and limitations (Gaudin & Chaliès, 2015). Such videos can be carefully chosen to highlight specific aspects of instruction and to allow program designers to build a set of stable resources surrounding each video—for example, mathematical explorations matched to video content, descriptions of the classroom context, and even videos showing contrasting methods for teaching identical content. This feature may result in more uniform professional development sessions. Stock video may also allow groups to engage in more critical discourse. For example, researchers found that teachers became more involved in discussions about video-recorded lessons of unknown teachers in comparison with teachers who watched one another's lessons (Seidel et al., 2011). Also, teachers who watched unknown teachers' videos were more reflective and discussed the consequences of the teachers' actions in greater depth (Kleinknecht & Schneider, 2013). However, stock video suffers from the problem of distance from teachers' classrooms; in particular, teachers who do not identify with the grade level, content taught, or the population of students may feel that the instructional practices highlighted by the video do not apply to them (Kleinknecht & Schneider, 2013; Seidel et al., 2011; Zhang et al., 2011).

Although empirical comparisons of these two conditions are scarce, one study that did make such comparisons found that having teachers watch their own lessons had greater positive effects by "providing teachers with a more activating experience" (Seidel et al., 2011, p. 266). Another study (Zhang et al., 2011) was more equivocal, finding benefits from both conditions. In viewing their own videos, teachers were able to see their classrooms from multiple angles. In addition, some teachers observed that their peers often noticed aspects of instruction that they themselves did not. However, some evidence in favor of stock video emerged as well; teachers who watched stock video found it to be a useful comparison with their own teaching and were able to modify their instruction by absorbing others' techniques (Zhang et al., 2011). Roth and colleagues (2011) used both stock- and own-video cases, in that order, in their successful program. Looking at researchers' insights regarding this progression, we extend this line of research by contrasting two conditions: one in which teachers watch only stock video and one in which teachers progress from stock video to watching and analyzing their own videos.

Type of Facilitation

In the last two decades, professional development models that are more teacher-directed, such as teacher inquiry groups, lesson study, and professional learning communities, have proliferated in schools alongside more traditional, facilitator-directed models. Advocates suggest that teacher-directed groups can take an inquiry stance regarding their own practice and collectively work to address dilemmas in teaching and learning (Cochran-Smith & Lytle, 2009; Garet et al., 2001; Hargreaves, 1996; Loucks-Horsley, 1996; MacDonald & Shirley, 2009; Stoll, Bolam, McMahon, Wallace, & Thomas, 2006). In these settings, teachers and facilitators co-construct knowledge; neither holds ultimate authority and teachers generally command a strong role in the development of craft knowledge and best practice (Cochran-Smith & Lytle, 2009). Such teacher-directed groups can be advantageous both politically, as they make clear that teachers hold the knowledge to improve instruction, and practically, as such groups do not require costly external facilitation.

Despite interest in teacher-directed professional development, few studies have examined the effect of external versus teacher facilitation on program processes and outcomes (Coles, 2014). In the area of video-based professional development, Calandra and colleagues (2009) observed nonfacilitated, preservice teacher video-based reflection sessions and noted that teachers attended to superficial features of practice, such as talking speed or time management. van Es and Sherin (2006) found that discussions in a teacher-led video group stayed broad and addressed a wide range of topics, such as classroom climate and management. By contrast, results from studies in which facilitators led group discussions showed a tighter focus on salient features of mathematics practice (Borko et al., 2008). They further found that a facilitated group developed a more focused and interpretive view of student thinking when analyzing videos of mathematics instruction. Borko and colleagues (2008) noted that facilitators "provided more specific guidance for the discussions around video and encouraged the teachers to take a more critical stance" (p. 433).

Of the above, only one study (van Es & Sherin, 2006) directly compared facilitated and nonfacilitated groups while holding the content of the professional development constant. Furthermore, a literature search yielded no studies comparing *types* of facilitation (e.g., teacher-led or facilitator-led). Sessions that are teacher-led but that have an external facilitator present may combine the attractive features of both levels. Given that even advocates of teacher-directed professional development generally acknowledge a role for a facilitator (Richardson, 2003), the present study varied the types of facilitation during meetings to help determine how this variable influences program outcomes.

The Present Study

With these issues in mind, we aimed to design teacher professional development focused around the analysis of video with a specific observational tool—the *Mathematical Quality of Instruction* instrument (MQI). The MQI was developed to focus on and capture the quality of mathematics in elementary classrooms offered by teachers and/or students (Hill et al., 2008). It contains four main dimensions with multiple codes beneath each (see Table 1). The instrument was developed from a grounded analysis of classroom video (Hill et al., 2008). This analysis revealed that teachers varied in the amount of meaning, sense-making, and disciplinary practices offered to students (*Richness of the Mathematics*); in the extent to which teachers took up student ideas or remediated student misconceptions (*Working With Students and Mathematics*); in the presence of teacher errors (*Errors and Imprecision*); and in the extent to which students engage in mathematical reasoning, explanation, and high-cognitive demand tasks (*Common Core Aligned Student Practice*). Similar to the argument made in Hiebert and Grouws (2007), this grounded analysis found that these dimensions varied independently. For instance, teachers who offered students opportunities to work on high-cognitive demand tasks did not always have a strong meaning orientation, and those with a strong meaning orientation sometimes made mathematical errors during instruction.

Each code on the MQI has a description and is scored as either not present, low, mid, or high. As an example, the *Explanations and Mathematical Sense-Making* code aims to capture instances of why—why a mathematical procedure works, why a solution method is appropriate, or why an answer is true. If no mathematical explanations or instances of sense-making are present, the clip is scored as not present for this code. A clip from a lesson is scored as low if there is an isolated explanation or instance of sense-making. It is scored mid if two or more brief explanations are present in

Table 1. MQI Dimensions and Codes.

Richness of the Mathematics	Working With Students and Mathematics	Errors and Imprecision	Common Core Aligned Student Practices
• Linking between representations • Explanations • Mathematical sense-making • Multiple procedures or solution methods • Patterns and generalizations • Mathematical language	• Remediation of student errors and difficulties • Teacher uses student mathematical contributions	• Mathematical content errors • Imprecision in language and notation • Lack of clarity in presentation of mathematical content	• Students provide explanations • Student mathematical questioning and reasoning • Students communicate about the mathematics • Task cognitive demand • Students work with contextualized problems

Note. MQI = Mathematical Quality of Instruction instrument.

the clip or an explanation is more than briefly present. Finally, a clip is scored as high if one or more explanation(s)/attention to sense-making is the focus of instruction and teacher–student discourse. The MQI draws observers' attention to these mathematics-specific elements of instruction and away from other classroom issues.

Researchers have recently connected teachers' MQI scores and student outcomes. Specifically, Hill, Umland, and Kapitula (2011) found that the MQI's overall measure of mathematics instructional quality also correlated with student outcomes. Blazar (2015) found that lessons that received high scores in the grouping of the three dimensions *Richness of the Mathematics, Working With Students and Mathematics*, and *Common Core Aligned Student Practices* were positively correlated with student outcomes, whereas the dimension *Errors, Imprecision*, and *Lack of Clarity* had a negative correlation with student outcomes. Garet and colleagues (2016) found that the dimension *Errors and Imprecision* was predictive of student performance (e.g., more errors, lower student performance).

The research literature suggests that using such an observation tool for professional development could provide teachers concrete opportunities to reflect on and improve their practice. Specifically, researchers have noted that without a lens or framework to view instruction, teachers might not notice what researchers intend for them to notice (Star & Strickland, 2008) or they may develop what Mason (2011) referred to as "fragmented awareness" with "inconsistent orientations" (Roth-McDuffie et al., 2013, p. 248). In contrast, researchers have found that frameworks for viewing lessons provide structure, focus, and a common orientation (Roth-McDuffie et al., 2013; Walkoe, 2015). The objective of using the MQI in professional development settings is not to classify mathematics instruction as good or bad; instead the goal is to more deeply understand the quality of mathematics found in lessons and to connect that understanding to teaching practices.

Prior to full implementation, we piloted MQI-based professional development with a small group of teachers and revised the instrument based on their feedback. For example, in the *Errors and Imprecision* dimension, teachers disliked being scored on three different types of errors (major mathematical errors, imprecision, and lack of clarity). Thus, those codes were collapsed into one. The result of this piloting is the *Mathematical Quality of Instruction for Professional Development* instrument (MQI-PD), which consists of a subset of dimensions and items from the full-length MQI (Hill et al., 2008).

Once we finalized the MQI-PD, we next used a small, exploratory study to determine which condition, or combination of conditions, would be best for taking into a more formal, rigorous field trial. Because we expected that video source and facilitation type might interact to produce unique outcomes, we selected a factorial design (Montgomery, 1991). Similar to other professional development studies (Heller, Daehler, Wong, Shinohara, & Miratrix, 2012), we held the content of the professional development—in this case, the MQI-PD and general structure of the professional development sessions—constant across treatment groups to allow an independent estimate of the role of video source and facilitation type. Because outcomes may vary by local context and facilitator, we recruited three groups for each condition, for a total of 12 treatment groups, to understand the effects of each condition.

We hypothesized that participation in this professional development and/or in specific conditions would lead to differences in the way teachers thought about and analyzed instruction. Specifically, we asked about professional development *processes*, investigating whether video source or facilitation type influences the quality of teachers' conversations. We also explored our process data for the potential emergence of a "culture of nice" in the own-video condition. Then we asked three questions about *outcomes*:

1. Does participation in the professional development program lead teachers to use concepts and knowledge from the MQI-PD to structure their analysis of practice, as contrasted with teachers in a comparison group?
2. What role does the *type of video* or *facilitation* play in teachers' use of concepts and knowledge from the MQI-PD to structure their analysis of practice?

Table 2. District Participation.

District	Description	% minority	% FRPL	Year 1 condition(s)	Year 2 condition(s)
1	Large urban	85.8	80	• Facilitator-led, own-video • Teacher-led, stock-video	• Teacher-led, own-video
2	Mid-sized urban	60.2	43	• Facilitator-led, stock-video • Teacher-led, own-video	NA
3	Small urban	42.2	40	NA	• Facilitator-led, own-video • Teacher-led, stock-video
4	Suburban	25.4	12	NA	• Facilitator-led, stock-video
5	Suburban	22	20	• Facilitator-led, stock-video • Teacher-led, own-video	• Teacher-led, stock-video
6	Mid-sized urban	30	48	NA	• Facilitator-led, own-video

Note. FRPL = free and reduced-price lunch.

3. Are participants more critical of their own instruction than teachers in a comparison group, and does facilitation type or video source play a role in encouraging teachers to become more critical?

The scope of this study did not allow us to examine instructional or leaning outcomes. Based on the research literature (e.g., Allen, Pianta, Gregory, Mikami, & Lun, 2011; Borko et al., 2010; Christ et al., 2014; Cohen, 2004; Kersting et al., 2010; Kersting et al., 2012; Kiemer, Gröschner, Pehmer, & Seidel, 2014; Seago et al., 2014), we reasoned that changes in teacher thinking are a first step toward changing classroom practice and improving student outcomes. For example, van Es, Tunney, Goldsmith, and Seago (2014) found that teachers learned to attend to students' mathematical thinking and as a result, changed their teaching practices to focus more on students' ideas (p. 341). Thus, we argue that using these proximal outcomes to test the intervention and its permutations is sufficient for exploratory, design-oriented trials like the one described here.

Method

We collected and analyzed qualitative and quantitative data to answer the research questions above, allowing us to triangulate between different sources of evidence regarding program and condition effectiveness in restructuring teachers' analyses of practice. We describe the specifics of our sample, intervention, data collection, and analyses below.

Sample

To recruit participants, we first approached leaders in local districts. In the districts whose leaders supported the study, the researchers gave a presentation about the study to teachers, and teachers voluntarily signed up for the study. Based on this approach, 126 treatment and 20 comparison group (untreated) teachers were recruited from six districts in the northeastern and northwestern United States. The teachers taught in Grades 3 through 8. While the original intent of the study was to randomly assign individual teachers to condition, we moved to assignment to group based on logistics (e.g., travel and schedule constraints, teachers' desire to be in the same group as their colleagues). As a result, 12 treatment groups were created by clustering teachers within location and, if there were enough teachers for multiple groups within a location, by schedule. For two sites, teachers were also clustered by grade level partly based on school release time for the different grade levels (see Table 2).

These groups were then each randomly assigned to a condition: own-video/facilitator-led, stock-video/facilitator-led, own-video/teacher-led, and stock-video/teacher-led. Teachers who initially enrolled but whose schedules did not permit them to attend the professional development formed a comparison group. Of the original 146 teachers, 26 withdrew before fall treatment sessions, citing reasons such as scheduling issues and job changes, four withdrew during the study for personal or medical reasons, and 14 completed the initial and/or midpoint survey but did not complete the final measures. The program operated across two academic years, with a small number of comparison group teachers becoming treatment group teachers in the second year of the study. This design is similar to others in the professional development field (see, for example, Roschelle et al., 2010; Sample-McMeeking, Orsi, & Cobb, 2012).

Procedure

The MQI-PD program consisted of three phases: an initial, standardized training in which all treatment teachers participated, regardless of condition; a 4-week period in which six groups experienced facilitator-led sessions and the other six experienced teacher-led sessions; and a 6-week period in which half of the treatment groups (i.e., half each of the facilitator-led and teacher-led groups) were invited to use their own video during the sessions (see Figure 1).

We delayed the start of the own-video sessions until the fifth week to allow teachers to become comfortable with one another, the facilitator, the instrument, and discussions about instruction. Six facilitators were assigned to these groups

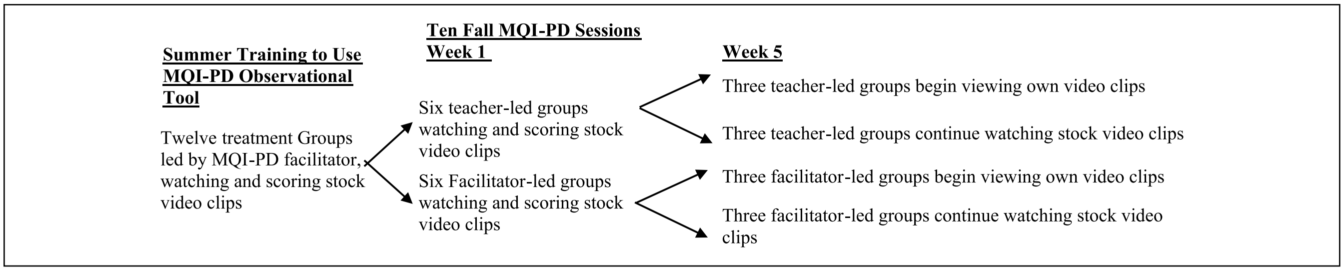

Figure 1. MQI-PD Treatment Groups.

based on geographical proximity and with the goal of having each facilitator work in more than one condition over the course of the study.[1] We describe each phase of the professional development program in turn.

An initial standardized training session was delivered to all treatment groups over 15 hr before the beginning of the school year to familiarize teachers with the dimensions of the MQI-PD. Because all groups received the same training, its delivery did not affect our ability to answer the research questions above. Facilitators introduced each MQI-PD code and its scoring rules to participants, then participants viewed video clips of diverse classroom instruction (between 3 and 5 min in length) and assigned tentative scores. Participants were provided transcripts for each of the video clips so they could ground their scoring rationales in evidence from the clip. Facilitators modeled scoring conversations for the participants and provided guidelines for discussing videos (Borko et al., 2008; Zhang et al., 2015; Zhang et al., 2011), with an emphasis on using evidence from the video clips and language from the MQI-PD for discussing and justifying scores (Borko et al., 2010; Goldsmith & Seago, 2011). Teachers then scored the video individually and in small groups, receiving feedback from the facilitator on whether their reasoning matched that of expert MQI-PD users.

Weekly fall sessions began roughly 4 weeks after the summer training. Regardless of treatment, teachers solved a mathematics problem related to one of the video clips (Borko et al., 2010), then analyzed and scored two clips with a subset of MQI-PD items in these 2-hr sessions. The mathematics problems and stock-video clips were the same in all treatment groups for the first 4 weeks of the fall sessions, which researchers recommend as teachers learn the method of analyzing and discussing videos (Zhang et al., 2011). In the facilitator-led groups, facilitators were provided with scaffolds for leading the conversations, including specific questions to elicit evidence from the video clips and prompts to tie teachers' responses to the text of MQI-PD codes. In the teacher-led groups, participants were asked to volunteer to lead similar conversations, with the expectation that they would use the discussion models provided in the initial training. The external facilitators in teacher-led groups were instructed to remain silent unless the conversation went off topic for more than a few minutes.

From the fifth week on, participants in stock-video groups continued to watch clips from the MQI-PD video library. Participants in own-video groups voluntarily shared videos of their mathematics classrooms. Participants were asked to select a lesson to record and provided with the equipment to do so. They were also asked to select a 3- to 5-min clip for scoring and the MQI-PD dimensions their peers would use to analyze the clip. Teachers who shared a clip did not participate in its scoring conversation; instead, they were asked to comment and reflect at the conclusion of the conversation. Finally, if there were fewer than two clips volunteered by participants for a session, own-video-group facilitators supplemented the session with the same clips used in stock-video groups.

Data Collection and Analyses

Processes. To analyze professional development processes, we audio-recorded weekly sessions and analyzed a subset (Weeks 2, 4, and 9) for fidelity to the assigned condition, as well as to determine any differences in conversations depending on facilitation type and video type. We expected the structure of the sessions to be similar across all groups, but that the facilitator-led sessions would have more instances in which the facilitator pressed for score explanations or encouraged group participation. Finally, to determine whether the type of video had an impact on the participants' conversations, we looked at whether teachers in the own- or stock-video conditions were more likely to state the score and rationale they had assigned to a clip; whether they related to the teacher in the clip, students, or content in the video (e.g., making statements such as "When I teach fractions to fourth graders, I use a similar approach"); and if groups would come to consensus on how to score a clip. We expected that teachers in the own-video groups might be less inclined to talk in detail about their peers' clips, lest they be perceived as too critical, and that there would be occasions in which they would score their peers' clips with higher MQI-PD scores than warranted.

To explore whether teachers were less critical when discussing videos from colleagues' classrooms, we also created a metric flagging when teachers' scores were more lenient (higher) than master raters' scores (1) or more severe (lower) than master raters' scores (0). Master raters scored the participants' own-video clips specifically for this purpose. To account

Table 3. Lesson and Video Reflection Questions Pre-, Mid-, and Post-MQI-PD Participation.

	Own-lesson reflection questions
1	What were you hoping students would learn during this lesson?
2	What went well in this lesson?
3	Did you or your students provide any mathematical explanations during this lesson? If so, please provide an example of such an explanation.
4	How much did students engage in mathematical reasoning and inquiry during today's lesson? If students did engage in these practices, please give an example.
5	What, if anything, did not go well during the lesson?
6	Was there anything that you struggled with mathematically during this lesson? What was it? What made this difficult for you?
7	If you teach this lesson again next year, will you change anything? If so, what?
	Stock-video reflection questions
1	What stood out to you about the mathematics in the clip?
2	What stood out to you about the teaching in the clip?
3	What seems significant to you about this mathematics instruction?
4	In your view, is this teaching likely to lead to students learning the content? Why or why not?

Note. MQI-PD = Mathematical Quality of Instruction for Professional Development.

for truncated outcomes due to distributions of master scores (i.e., over-scoring automatically occurs when the master score is the lowest score point, and over-scoring never occurs when the master score is the highest score point), we only flagged instances where both over- and under-scoring were possibilities for teachers' incorrect responses. Using own-video teachers as our sample, we predicted teacher leniency using logistic regression with indicators for (a) whether the clip was derived from a video library or the classroom of a participating teacher, and (b) the teacher's measured leniency in the weeks prior to beginning the own-video condition.

Outcomes. To determine the effect of the professional development and/or specific conditions on the way teachers structured their analyses of practice, we examined teachers' open-ended responses to (a) short video clips and (b) reflections on their own lessons. Teachers responded to two clips and wrote two lesson reflections before, at the midpoint (Week 4), and at the end of the professional development. The comparison teachers were asked to respond to the same clips and write lesson reflections at the same time points as the treatment teachers. Table 3 shows the questions asked at these times.

We coded teachers' responses to determine the prevalence of MQI-PD-related considerations. This involved coding each response for the presence of *any* MQI-PD component addressed in the program (e.g., remediation of student errors, multiple solution methods, student engagement in mathematical practices), as well as language similar to that in the MQI-PD. When no MQI-PD component was mentioned, the response was labeled "no MQI-PD."

To assess the extent to which the MQI-PD honed teachers' definitions of commonly used mathematics education terms, we asked teachers whether students provided mathematical explanations and/or engaged in mathematical reasoning during the lessons they were reflecting upon. The MQI-PD defines mathematical explanations and reasoning with a very high bar; we were interested to know whether teachers had adopted that bar, and thus we asked them to provide a short description of that explanation and/or reasoning, then coded those examples to determine whether they met the MQI-PD criteria.

We expected participating teachers might become more critical of and reflective about their instruction, and thus our fifth and sixth questions asked about any struggles in their lessons. We characterized responses to the question as no struggles, struggles with content, struggles with delivery (e.g., helping students access the mathematics), or struggles with other (e.g., management, district demands). We also asked teachers whether they planned to make any changes to the lesson prior to teaching it again.

A team of authors and MQI-PD-trained graduate assistants used Dedoose to complete the coding described above. Coders were blind to the type of professional development each participant received and the teacher's district. When agreement on codes reached 75%, on average, we assigned pairs of coders batches of reflections to code independently. Coders were asked to reconcile any discrepancies in assigned codes. One author triple-coded a random selection of about 45% of the reflections to assess coding accuracy. Our theory of action held that if successful, the MQI-PD would change how teachers analyzed both the teaching of others and their own practice; how teachers use mathematical language like "explanation" and "reasoning" in ways that are consistent with their MQI-PD definitions; and that teachers would become more likely to critically reflect on and want to change their own practice, particularly around their delivery of content, which is the major focus of the MQI-PD. Thus, after coding was completed, we created seven distinct measures at each time point to capture these outcomes:

- Teacher made MQI-PD-related comments in the video clip reflections
- Teacher made MQI-PD-related comments in own-lesson reflections
- Teacher identified a student explanation that met the MQI-PD definition
- Teacher identified an instance of student reasoning that met the MQI-PD definition
- Teacher reported one or more general struggles
- Teacher reported at least one struggle with the delivery of content

Table 4. Video Clip Sharing in Own-Video Groups.

	Year 1				Year 2			
District	Format	Teachers	Clips shared	Description	Condition	Teachers	Clips shared	Description
1	Facilitator-led	8	3	3 teachers, 1 clip each	Teacher-led	9	4	4 teachers, 1 clip each
2	Teacher-led	6	9	3 teachers, 2 clips each; 3 teachers, 1 clip each	Stock-video	—	—	—
3	Stock-video	—	—	—	Facilitator-led	14	8	6 teachers, 1 clip each; 1 teacher, 2 clips
4	Stock-video	—	—	—	Stock-video	—	—	—
5	Teacher-led	7	12	2 teachers, 1 clip each; 5 teachers, 2 clips each	(none shared)			
6	Stock-video	—	—	—	Facilitator-led	5	8	4 teachers, 2 clips each

- Teacher planned to make changes in future lessons (coded categorically: 0 items, 1 item, 2-3 items, >3 items)

We conducted two types of analyses on each of these measures. We first used Poisson regression models with treatment as a dummy variable to determine the overall main effect of attending one or more MQI-PD professional development sessions (our treatment group) versus remaining in the comparison group. Second, because treatment teachers were further grouped into different conditions, we conducted additional Poisson regression analyses controlling for these conditions, to investigate whether there existed a heterogeneous effect of MQI-PD across conditions.[2] Unless otherwise noted, measures have been kept in their original scale, and we used a relaxed definition of statistical significance ($p < .10$) because of small sample size, because of our ability to triangulate between qualitative and quantitative data, and because the risk in incorrectly accepting the null hypothesis (i.e., failing to move the program forward or to take up a specific design) is greater than the risk of falsely rejecting the null hypothesis (Shavelson, 1996).

Results

Processes

Does the quality of teachers' conversations vary by facilitation type or video source? We found that across all four conditions, most sessions followed the intended structure, with only minor differences among facilitators in pacing and order. However, sessions did differ in process according to condition. A review of session transcripts suggests that most facilitator-led group leaders maintained their groups' focus on the video clip and pressed participants to share their scores and rationales with the group. Facilitators also addressed misconceptions about the MQI-PD items and kept participants involved in the conversations. In teacher-led groups, the volunteer facilitators made many fewer of the scripted facilitator moves, played a minimal role in moving the conversation along, and were less likely to address misconceptions about the MQI-PD items. Still, teachers in these groups referred to the language of the MQI-PD, cited evidence in the clip, and expressed different opinions as often as teachers in the facilitator-led groups.

In the own-video groups, there was significant variation in the extent to which participants shared video clips from their classrooms. Two of the six groups had full teacher participation, one group had four out of five teachers share video clips, and other groups had minimal participation (see Table 4).

Based on the literature described above, we expected that teachers in the own-video groups would be less likely to state their rationales, and more likely to make statements relating to their peers' videos and to come to agreement about how they scored their peers' instruction. However, we found only one discernible difference among the groups. In particular, we found teachers in the own-video groups more likely to come to consensus about MQI-PD scores on their peers' clips, with more willingness to agree to disagree about the clips in the stock-video groups.

Did teachers assign higher-than-warranted MQI-PD scores to their peers' lessons? Next, we explore whether teachers were more lenient (i.e., assigned higher scores) when discussing videos made by their colleagues as compared with stock videos. For this analysis, we used a logistic regression to predict whether own-video teachers "over-scored" clips recorded by teachers in the group. Our analyses suggest that this was the case ($b = 1.28$, $p < .00$). This effect appeared even while controlling for teachers' initial propensity to over-score (i.e., controlling for proportion of over-scores in Weeks 1-4) as well as the number of opportunities for over-scoring.

Outcomes

Does participation lead teachers to use concepts and knowledge from the MQI-PD to structure their analysis of practice, as contrasted with teachers in a comparison group? Table 5 displays our models relating teachers' analysis of video to their participation in the professional development, controlling for their initial video analysis score.

Participants in the professional development were significantly more likely to notice MQI-PD-related topics in the videos shown at the midpoint and the endpoint of the study (all models).[3] Thus, teachers participating in the program used the MQI-PD to guide their analysis of video-recorded instruction, as compared with teachers in the comparison group. However, this treatment effect did not carry over into teachers' reflections on their own lessons; in Table 6, we see that, while generally positive, the treatment was not significantly associated with a stronger use of MQI-PD considerations in teachers' analyses of their own lessons.

Next, we turn to evidence that treatment teachers incorporated MQI-PD definitions into their use of mathematical and pedagogical terminology. Specifically, we had expected the professional development would help align teachers' examples of student mathematical explanations and reasoning, elicited in their lesson reflections, and the definitions the MQI-PD uses for these terms. Coding for these items indicated when teachers' examples met the MQI-PD bar, and thus Tables 7 and 8 show the relationship between professional development participation and midpoint and endpoint definitional agreement (columns 1 and 3). Participation in the program was unrelated to teachers' examples of students' classroom explanations meeting the MQI-PD definitional bar; the same was true for the student reasoning variable.

Does type of video or type of facilitation play a role in teachers' use of concepts and knowledge from the MQI-PD to structure their analysis of practice? Next, we look at whether the type of facilitation or origin of the video was associated with teachers' use of the MQI-PD framework to analyze instruction. Our tables present results for each outcome in two stages. First, we depict effects on the midpoint assessment comparing facilitator-led and teacher-led groups (as all teachers to that point viewed stock video). Next, we examine the impact of facilitation type and video origin on our endpoint measures, by (a) combining across subgroups within type (e.g., facilitation type without distinguishing among origin of video) and (b) examining each cell individually, omitting the main treatment effect to aid interpretation (i.e., each group is compared with the comparison), and including F tests of differences between groups. The latter analysis may be problematic due to small sample sizes within each cell, but does provide evidence regarding the condition most conducive to our outcomes. ANCOVAs for these midpoint and endpoint outcomes show that, controlling for initial scores, sites explained between 5% and 14% of the variance in our outcomes.

Table 5 (columns 2, 4, 5, and 6) shows that no condition had an advantage over the others on teachers' use of the MQI-PD to analyze stock-video clips. Thus, while the overall treatment was associated with teachers' use of the MQI-PD in analyzing elements of video clips, there was no design element that in particular worked to foster this analysis. This is represented by the p values of the F tests comparing the effects of each condition to the others. For lesson reflections (see Table 6), we found that analyzing own videos in teacher-led groups did predict an increase in use of MQI-PD categories in teachers' endpoint responses when compared with responses from the comparison group. The comparisons of coefficients at the bottom of this table also suggest that outcomes for the teacher-led/own-video condition were also marginally more MQI-PD-related than the facilitator-led/stock-video (logit difference = 0.715, $p = .031$) and the teacher-led/stock-video conditions (logit difference = 0.734; $p = .059$). Although this analysis is tentative due to the sample size, this finding suggests that teachers' use of the MQI-PD to structure reflections on their own practice may best be enhanced through teacher-led sessions examining their own videos. It is worth remembering, however, that the own-video condition occurred after several sessions in which teachers analyzed stock video.

Columns 2, 4, 5, and 6 in Tables 7 and 8 show the effect of specific conditions on teachers' use of MQI-PD-aligned definitions for student explanations and student reasoning, respectively. There were no associations between participation in a specific condition and the provision of examples of student explanation that met the MQI-PD definition. For student reasoning, we observed a large negative effect of the facilitator-led condition overall, and the facilitator-led/stock-video condition as compared with the comparison group and the teacher-led conditions. However, the impact of the facilitator-led condition was positive and nonsignificant at the study's midpoint. Thus, overall, we conclude that there are no consistent patterns regarding treatment condition and teachers' deployment of MQI-PD-aligned definitions.

Are participants more critical of their own instruction as contrasted with teachers in a comparison group? Does facilitation type or video source play a role in encouraging teachers to become more critical? Tables 9, 10, and 11, columns 1 and 3, display our models of the relationship between treatment condition and teachers' reports of struggling with the lesson, struggling with delivery and content specifically when reporting a struggle, and their nominations of changes they would make when they subsequently reteach the lesson. The likelihood of a teacher reporting any struggle (Table 9) was not related to treatment condition, but the likelihood that a reported struggle would focus on the delivery of content (Table 10) was positive and statistically significant at the midpoint of the professional development, but not at the end. Teachers' reports of planned changes to their lessons were related to participation in treatment at the midpoint of the professional development but not at the end.

Tables 9, 10, and 11, columns 2, 4, 5, and 6, show the relationships between specific treatment conditions and teachers' reported struggles and changes. At the end of the program, teachers in the facilitator-led/stock-video condition

Table 5. Effects on MQI-PD-Related Comments—Stock-Video Reflection.

	Midpoint score		Endpoint score			
	1	2	3	4	5	6
Initial score	0.134*** (0.0448)	0.135*** (0.0454)	0.0729 (0.0533)	0.0737 (0.0533)	0.0716 (0.0523)	0.0708 (0.0509)
Treatment	0.565*** (0.157)	0.593*** (0.170)	0.772*** (0.201)	0.836*** (0.211)	0.747*** (0.206)	
Treatment conditions						
Facilitator-led		−0.0559 (0.110)		−0.131 (0.114)		
Own-video					0.0538 (0.115)	
Facilitator-Led × Own-Video						0.728*** (0.227)
Facilitator-Led × Stock-Video						0.676*** (0.213)
Teacher-Led × Own-Video						0.895*** (0.230)
Teacher-Led × Stock-Video						0.797*** (0.221)
Constant	0.454*** (0.175)	0.451** (0.176)	0.232 (0.218)	0.231 (0.218)	0.235 (0.217)	0.237 (0.216)
Observations	115	115	101	101	101	101
p value of *F* test comparing effects						
Facilitator-Led × Own-Video = Facilitator-Led × Stock-Video						.725
Facilitator-Led × Own-Video = Teacher-Led × Own-Video						.326
Facilitator-Led × Own-Video = Teacher-Led × Stock-Video						.675
Facilitator-Led × Stock-Video = Teacher-Led × Own-Video						.146
Facilitator-Led × Stock-Video = Teacher-Led × Stock-Video						.395
Teacher-Led × Own-Video = Teacher-Led × Stock-Video						.550

Note. Robust standard errors of estimates reported in parentheses. MQI-PD = Mathematical Quality of Instruction for Professional Development.
*$p < .1$. **$p < .05$. ***$p < .01$.

Table 6. Effects on MQI-PD-Related Comments—Own-Lesson Reflection.

	Midpoint score		Endpoint score			
	1	2	3	4	5	6
Initial score	0.0628 (0.119)	0.0484 (0.120)	0.349** (0.159)	0.372** (0.164)	0.350** (0.158)	0.323** (0.164)
Treatment	0.149 (0.258)	0.0559 (0.295)	0.261 (0.287)	0.370 (0.315)	0.0392 (0.312)	
Treatment conditions						
Facilitator-led		0.163 (0.221)		-0.188 (0.315)		
Own-video					0.453 (0.293)	
Facilitator-Led × Own-Video						0.252 (0.478)
Facilitator-Led × Stock-Video						0.0643 (0.328)
Teacher-Led × Own-Video						0.715** (0.320)
Teacher-Led × Stock-Video						−0.0189 (0.417)
Constant	−0.386 (0.264)	−0.369 (0.264)	−0.949*** (0.311)	−0.982*** (0.316)	−0.950*** (0.309)	−0.912*** (0.311)
Observations	118	118	103	103	103	103
p value of *F* test comparing effects						
Facilitator-Led × Own-Video = Facilitator-Led × Stock-Video						.686
Facilitator-Led × Own-Video = Teacher-Led × Own-Video						.316
Facilitator-Led × Own-Video = Teacher-Led × Stock-Video						.601
Facilitator-Led × Stock-Video = Teacher-Led × Own-Video						.0308
Facilitator-Led × Stock-Video = Teacher-Led × Stock-Video						.833
Teacher-Led × Own-Video = Teacher-Led × Stock-Video						.0592

Note. Robust standard errors of estimates reported in parentheses. MQI-PD = Mathematical Quality of Instruction for Professional Development.
$^{*}p < .1$. $^{**}p < .05$. $^{***}p < .01$.

Table 7. Effects on MQI-PD Alignment of Student Mathematical Explanation Description—Own-Lesson Reflection.

	Midpoint score		Endpoint score			
	1	2	3	4	5	6
Initial score	1.925* (1.112)	2.064* (1.120)	1.983* (1.179)	1.867 (1.171)	2.017* (1.186)	1.874 (1.165)
Treatment	−0.875 (0.680)	−1.064 (0.765)	0.668 (1.021)	0.783 (1.013)	0.449 (1.055)	
Treatment conditions						
Facilitator-led		0.408 (0.648)		−0.289 (0.527)		
Own-video					0.479 (0.519)	
Facilitator-Led × Own-Video						0.726 (1.111)
Facilitator-Led × Stock-Video						0.204 (1.178)
Teacher-Led × Own-Video						1.122 (1.100)
Teacher-Led × Stock-Video						0.566 (1.053)
Constant	−1.393** (0.643)	−1.423** (0.653)	−2.137** (1.037)	−2.105** (1.020)	−2.146** (1.044)	−2.107** (1.022)
Observations	117	117	102	102	102	102
p value of *F* test comparing effects						
Facilitator-Led × Own-Video = Facilitator-Led × Stock-Video						.520
Facilitator-Led × Own-VIDEO = Teacher-Led × Own-Video						.591
Facilitator-Led × Own-Video = Teacher-Led × Stock-Video						.814
Facilitator-Led × Stock-Video = Teacher-Led × Own-Video						.268
Facilitator-Led × Stock-Video = Teacher-Led × Stock-Video						.643
Teacher-Led × Own-Video = Teacher-Led × Stock-Video						.425

Note. Robust standard errors of estimates reported in parentheses. MQI-PD = Mathematical Quality of Instruction for Professional Development.
$^{*}p < .1$. $^{**}p < .05$. $^{***}p < .01$.

Table 8. Effects on MQI-PD Alignment of Student Mathematical Reasoning Description—Own-Lesson Reflection.

	Midpoint score		Endpoint score			
	1	2	3	4	5	6
Initial score	0.377 (0.808)	0.142 (0.823)	1.583* (0.896)	1.847** (0.922)	1.584* (0.893)	2.029** (0.981)
Treatment	0.387 (0.678)	−0.0339 (0.732)	−0.616 (0.656)	−0.126 (0.701)	−0.708 (0.697)	
Treatment conditions						
Facilitator-led		0.758 (0.471)		−1.075** (0.547)		
Own-video					0.202 (0.509)	
Facilitator-Led × Own-Video						−0.716 (0.813)
Facilitator-Led × Stock-Video						−1.938** (0.975)
Teacher-Led × Own-Video						−0.188 (0.835)
Teacher-Led × Stock-Video						−0.0645 (0.753)
Constant	−1.464** (0.648)	−1.415** (0.648)	−0.790 (0.634)	−0.846 (0.648)	−0.790 (0.634)	−0.885 (0.661)
Observations	117	117	102	102	102	102
p value of *F* test comparing effects						
Facilitator-Led × Own-Video = Facilitator-Led × Stock-Video						.185
Facilitator-Led × Own-Video = Teacher-Led × Own-Video						.488
Facilitator-Led × Own-Video = Teacher-Led × Stock-Video						.330
Facilitator-Led × Stock-Video = Teacher-Led × Own-Video						.0643
Facilitator-Led × Stock-Video = Teacher-Led × Stock-Video						.0333
Teacher-Led × Own-Video = Teacher-Led × Stock-Video						.857

Note. Robust standard errors of estimates reported in parentheses. MQI-PD = Mathematical Quality of Instruction for Professional Development.
$^{*}p < .1$. $^{**}p < .05$. $^{***}p < .01$.

Table 9. Effects on Teacher-Reported Struggles—Own-Lesson Reflection.

	Midpoint score		Endpoint score			
	1	2	3	4	5	6
Initial score	0.998** (0.408)	1.008** (0.410)	0.744* (0.436)	0.753* (0.438)	0.767* (0.446)	0.744* (0.448)
Treatment	0.143 (0.559)	−0.0681 (0.591)	0.627 (0.598)	0.391 (0.634)	0.679 (0.634)	
Treatment conditions						
Facilitator-led		0.449 (0.436)		0.484 (0.446)		
Teacher-led					−0.117 (0.455)	
Facilitator-Led × Own-Video						0.393 (0.688)
Facilitator-Led × Stock-Video						1.461* (0.790)
Teacher-Led × Own-Video						0.823 (0.799)
Teacher-Led × Stock-Video						0.168 (0.671)
Constant	−0.173 (0.569)	−0.178 (0.569)	−0.676 (0.632)	−0.683 (0.633)	−0.693 (0.636)	−0.676 (0.635)
Observations	118	118	103	103	103	103
p value of *F* test comparing effects						
Facilitator-Led × Own-Video = Facilitator-Led × Stock-Video						.124
Facilitator-Led × Own-Video = Teacher-Led × Own-Video						.540
Facilitator-Led × Own-Video = Teacher-Led × Stock-Video						.685
Facilitator-Led × Stock-Video = Teacher-Led × Own-Video						.434
Facilitator-Led × Stock-Video = Teacher-Led × Stock-Video						.0516
Teacher-Led × Own-Video = Teacher-Led × Stock-Video						.342

Note. Robust standard errors of estimates reported in parentheses.
$^{*}p < .1$. $^{**}p < .05$. $^{***}p < .01$.

Table 10. Effects on Teacher-Reported Delivery/Content Struggles—Own-Lesson Reflection.

	Midpoint score		Endpoint score			
	1	2	3	4	5	6
Initial score	0.844 (0.561)	0.798 (0.553)	0.0996 (0.603)	0.153 (0.599)	0.109 (0.606)	0.127 (0.633)
Treatment	1.675** (0.702)	1.979** (0.825)	−0.864 (1.184)	−1.011 (1.196)	−0.833 (1.227)	
Treatment conditions						
Facilitator-led		−0.547 (0.636)		0.300 (0.576)		
Own-video					−0.0653 (0.588)	
Facilitator-Led × Own-Video						−0.377 (1.305)
Facilitator-Led × Stock-Video						−0.950 (1.300)
Teacher-Led × Own-Video						−1.414 (1.270)
Teacher-Led × Stock-Video						−0.669 (1.284)
Constant	−0.798 (0.703)	−0.766 (0.700)	1.527 (1.334)	1.483 (1.326)	1.519 (1.336)	1.504 (1.350)
Observations	76	76	61	61	61	61
p value of *F* test comparing effects						
Facilitator-Led × Own-Video = Facilitator-Led × Stock-Video						.503
Facilitator-Led × Own-Video = Teacher-Led × Own-Video						.248
Facilitator-Led × Own-Video = Teacher-Led × Stock-Video						.743
Facilitator-Led × Stock-Video = Teacher-Led × Own-Video						.572
Facilitator-Led × Stock-Video = Teacher-Led × Stock-Video						.722
Teacher-Led × Own-Video = Teacher-Led × Stock-Video						.390

Note. Robust standard errors of estimates reported in parentheses. Sample includes teachers who reported a struggle.
*$p < .1$. **$p < .05$. ***$p < .01$.

Table 11. Effects on Teacher-Reported Desired Changes—Own-Lesson Reflection.

	Midpoint score		Endpoint score			
	1	2	3	4	5	6
Initial score	1.188*** (0.373)	1.165*** (0.367)	0.904** (0.451)	0.963** (0.472)	0.907** (0.452)	0.990** (0.417)
Treatment	1.296* (0.667)	1.516** (0.693)	−0.0781 (0.721)	−0.364 (0.767)	−0.0976 (0.745)	
Treatment conditions						
Facilitator-led		−0.434 (0.362)		0.545 (0.380)		
Own-video					0.0432 (0.377)	
Facilitator-Led × Own-Video						0.388 (0.689)
Facilitator-Led × Stock-Video						−0.0437 (0.698)
Teacher-Led × Own-Video						−0.711 (0.737)
Teacher-Led × Stock-Video						−0.167 (0.677)
Constant cut 1	−1.498** (0.737)	−1.527** (0.737)	−1.514* (0.835)	−1.482* (0.846)	−1.511* (0.836)	−1.483** (0.735)
Constant cut 2	0.931 (0.731)	0.905 (0.732)	−0.209 (0.781)	−0.162 (0.794)	−0.206 (0.782)	−0.150 (0.702)
Constant cut 3	3.241*** (0.829)	3.227*** (0.828)	2.297*** (0.832)	2.372*** (0.848)	2.301*** (0.832)	2.415*** (0.748)
Constant cut 4	5.290*** (0.895)	5.293*** (0.893)	4.223*** (1.000)	4.302*** (1.003)	4.226*** (1.000)	4.353*** (0.880)
Observations	118	118	103	103	103	103
p value of *F* test comparing effects						
Facilitator-Led × Own-Video = Facilitator-Led × Stock-Video						.440
Facilitator-Led × Own-Video = Teacher-Led × Own-Video						.0718
Facilitator-Led × Own-Video = Teacher-Led × Stock-Video						.298
Facilitator-Led × Stock-Video = Teacher-Led × Own-Video						.274
Facilitator-Led × Stock-Video = Teacher-Led × Stock-Video						.818
Teacher-Led × Own-Video = Teacher-Led × Stock-Video						.352

Note. Robust standard errors of estimates reported in parentheses.
*$p < .1$. **$p < .05$. ***$p < .01$.

were significantly more likely to report a struggle ($p < .10$) than the comparison group (Table 9). An F test shows that the difference between facilitator-led/stock-video teachers to teacher-led/stock-video teachers in the number of struggles reported in their own-lesson reflections at the endpoint was marginally significant.

For treatment and comparison teachers who did report some struggle in their lesson reflections ($n = 76$ at the midpoint; $n = 61$ at the endpoint), Table 10 did indicate a treatment effect (against our comparison group) on whether the struggle reported was related to the delivery of the mathematics or the mathematical content of the lesson. This effect disappeared, however, at the endpoint, and no specific condition appeared to be more likely to report such struggles, rendering the overall interpretation inconclusive.

Discussion

This study explored whether teachers' participation in MQI-PD and, in particular, four conditions formed by the permuting of facilitator/teacher-led and stock/own-videos led to changes in teacher analysis of instruction, language use, and critical reflection on their own practice. Based on the research literature (e.g., Allen et al., 2011; Borko et al., 2010; Christ et al., 2014; Cohen, 2004; Kersting et al., 2010; Kersting et al., 2012; Kiemer et al., 2014; Seago et al., 2014), we reasoned that such changes are a first step toward changing classroom practice and improving student outcomes. We also investigated whether any of the four conditions was related to differences in the nature of MQI-PD conversations.

The results illustrate that teachers had similar depth and focus in their conversations regardless of the type of facilitation or type of video. Although the literature suggests that teacher-led groups struggle with maintaining focus and have more superficial discussions (Calandra, Gurvitch, & Lund, 2008; van Es & Sherin, 2006), we found no significant difference in the quality of conversations between teacher-led and facilitator-led groups. Similar to the facilitator-led groups, the teacher-led groups often worked through dissenting views and pressed each other for explanations and more information. This suggests that teachers can utilize the MQI-PD procedures in video clubs of their own.

That said, we note that all groups of teachers were provided with the same 15-hr training on the MQI-PD, during which facilitators demonstrated conversation norms (Borko et al., 2008; Sherin & van Es, 2005; van Es et al., 2014). External facilitators were also present in teacher-led groups, further reinforcing those norms. We believe that these norms prompted teacher-led groups to have conversations similar to those of the facilitator-led groups. This is in contrast to other studies in which teachers were not provided with a particular focus with which to view and talk about mathematics instruction. In these cases, teachers' conversations stayed broad and addressed a wide range of topics (van Es & Sherin, 2006).

In the own-video groups, fewer teachers volunteered to share videos than we had expected and there was significant variation among the groups in this regard (Desimone & Garet, 2015). We noted that groups with more sharing of videos tended to come from districts and schools with a history of close collaboration and public discussion of instruction, which Borko and colleagues (2008) noted are important for teachers to feel comfortable sharing video clips. In addition, teachers who scored peers' videos were more likely to assign generous scores. This suggests that teachers encouraged a "culture of nice" through the scores they assigned—as the scores assigned to own-video clips were quite high on average, teachers avoided more difficult conversations regarding what was missing from instruction or how it could be improved (MacDonald, 2011). Both these findings suggest that the own-video program, as designed, would be best suited to contexts in which strong collaborative and critical norms exist. Alternatively, we could strengthen the program to develop such norms or recruit groups of teachers based on a willingness to engage in public sharing.

When examining teachers' analysis of video clips at the study mid and end time points, we found that teachers in the professional development groups were more likely than teachers in the comparison group to reference MQI-PD-related topics. However, this MQI-PD noticing did not carry over to reflections on their own lessons, with the exception of teachers who were in the teacher-led, own-video groups. Our results confirm Zhang et al.'s (2011) conclusion that both conditions—teachers' own and stock video—could benefit teachers.

With regard to whether participants used MQI-PD-compatible definitions of student mathematical explanations and reasoning, we found no patterns, either in relationship to the comparison group or among the four treatment conditions. More generally, a scan of teachers' lesson reflections at the end of the study suggested that many—perhaps most—of the participants' comments focused on whether students "got it or not." Participants focused on students' content misunderstandings (e.g., some students struggled with multiplication and division facts) or process difficulties (e.g., not all students participated in the whole group discussion, some were still hesitant about joining in on the shared thinking), followed by management issues, mostly around time.

Another question was whether the treatment generally or a specific condition would lead teachers to become more critical of their own instruction. At the conclusion of the study, we found that teachers in the facilitator-led, stock-video groups were more likely to report a struggle in a recently taught lesson than the comparison group. This finding aligns with Seidel et al.'s (2011) conclusion that teachers in the stock-video condition were more critical and found more alternatives than teachers in the own-video condition, and that teachers in the own-video condition were "less reflective when articulating critical incidents" (p. 259).

Teachers' noticing of the MQI-PD-related aspects of their lessons may be due to the different types of access teachers had to the stock-video clips as compared with their own lessons that they reflected on. Teachers could pause, rewind, and re-watch the stock videos, effectively providing them with more opportunities to look for concrete evidence and to cite specific instances of instruction related to the MQI-PD. In contrast, teachers did not have the opportunity to revisit exactly what had occurred in their own lessons, relying on their memories of what had happened. However, teachers in the own-video groups had the experience of watching and noticing MQI-PD features of their own and their peers' lessons, leading to a greater ability to recall MQI-PD features of their past lessons. Not having the ability to revisit their lessons also may have affected teachers' facility in pointing out examples of students' explanations and reasoning in the stock clips and their lessons.

Conclusion

The design of the MQI-PD program incorporated what are considered to be the best practices in professional development, such as a focus on mathematical content, active learning, lengthy duration, and collective participation (Borko, 2004; Desimone, 2009; Desimone & Garet, 2015). Recently, Desimone and Garet (2015) called for more evidence of "specific, effective activities" that would provide additional insight into the efficacy of features of professional development programs (p. 260) and our study investigated the efficacy of the type of video and type of facilitation in the context of the MQI-PD program. Based on our findings, we believe we have some evidence, both from group conversations and teachers' video analyses and lesson reflections, that the teacher-led, own-video condition is slightly superior to all other conditions. This finding confirms Desimone and Garet's (2015) conclusion that professional development has greater success when it is directly connected to teachers' lessons (p. 254). Even though they were generous to peers in their scores, teachers in this teacher-led, own-video condition were just as likely to have high-quality conversations and were more likely to use the MQI-PD in video analysis. That said, our results suggest that efforts intended to enact this kind of professional development may take a significant amount of time to cultivate—even though teachers had 15 hr of MQI-PD summer training and four 2-hr sessions in which teachers discussed videos of instruction, many teachers still did not feel comfortable enough to share videos from their own classrooms.

While we observed positive results in the teacher-led, own-video condition, the MQI-PD intervention had limited impact on teachers' reflections on their own practices, which we considered a first step toward changing classroom practice and improving student outcomes. Still, our findings indicate that introducing the professional development program through facilitator-led conversations (to model and establish discussion norms) and stock video (to gain experience watching and scoring lessons) are critical features that prompted teachers to engage in meaningful discussions about mathematics instruction. We recommend that teachers gain experience with the MQI-PD (or possibly another professional development program) before transitioning to the teacher-led, own-video condition.

We also believe, based on teachers' reflections on their own instruction and the mixed results, that embedding professional development in practice may not be enough; there may need to be opportunities to put strategies into practice and receive feedback. We posit that a stronger intervention with some accountability for instructional change is necessary (Tripp & Rich, 2012). This program improved teachers' ability to analyze video clips, but the kinds of thinking deployed in analyzing clips did not transfer to thinking about their own lessons, perhaps because we did not promote or emphasize that transfer as part of the professional development. Schön (1983) noted that reflecting on practice should lead a teacher to action. Thus, we suggest that a clear, concrete bridge that leads teachers to explicitly apply video-watching protocols to their own practice may be essential for success.

There are limitations to the study described in this article. First, the number of teachers per group and the number of groups per condition were relatively small. Without a larger sample of groups or random assignment of individual teachers to condition, we cannot rule out the possibility that our results were driven by the composition of particular groups rather than the professional development treatments. Second, despite our best efforts, not all teachers in own-video groups shared clips of their instruction. It is possible that more skillful facilitation, or more collaborative contexts, may have enabled more sharing of video, and thus more impact on teachers' thinking about their own instruction. Third, in retrospect and in keeping with our reading the above literature, we would design a stronger intervention, one that asked teachers, for instance, to regularly use MQI-PD principles to reflect on their own instruction. Learning to analyze others' classroom video may not necessarily affect teachers' analysis of their own. As such, this limitation suggests to the teacher education community program features that must be present in video-based professional development.

Declaration of Conflicting Interests

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

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research reported here was supported by the National Science Foundation through Grant 1221693 to Harvard University. The opinions expressed are those of the authors and do not represent views of the National Science Foundation.

Notes

1. Two facilitators led only one group.
2. For the outcomes assessing whether teachers' examples of student explanations or reasoning met *Mathematical Quality of Instruction for Professional Development* instrument (MQI-PD) standards, and for the outcomes considering teachers' reported struggles, we used logistic regression modeling. For the outcome capturing categorically the number of changes teachers reported planning to make in future lessons, we used ordered logistic regression.
3. Poisson regression models the log of the expected count outcome on predictors; coefficients should be interpreted thusly.

References

Allen, J. P., Pianta, R. C., Gregory, A., Mikami, Y., & Lun, J. (2011). An interaction based approach to enhancing secondary school instruction and student achievement. *Science*, *333*(6045), 1034-1037.

Blazar, D. (2015). Effective teaching in elementary mathematics: Identifying classroom practices that support student achievement. *Economics of Education Review*, *48*, 16-19.

Borko, H. (2004). Professional development and teacher learning: Mapping the terrain. *Educational Researcher*, *33*(8), 3-15.

Borko, H., Jacobs, J., Eiteljorg, E., & Pittman, M. E. (2008). Video as a tool for fostering productive discussions in mathematics professional development. *Teaching and Teacher Education*, *24*(2), 417-436.

Borko, H., Koellner, K., Jacobs, J., & Seago, N. (2010). Using video representations of teaching in practice-based professional development programs. *ZDM*, *43*(1), 175-187.

Calandra, B., Brantley-Dias, L., Lee, J. K., & Fox, D. L. (2009). Using video editing to cultivate novice teachers' practice. *Journal of Research on Technology in Education*, *42*(1), 73-94.

Calandra, B., Gurvitch, R., & Lund, J. (2008). An exploratory study of digital video editing as a tool for teacher preparation. *Journal of Technology and Teacher Education*, *16*(2), 137-153.

Christ, T., Arya, P., & Chiu, M. M. (2014). Teachers' reports of learning and application to pedagogy based on engagement in peer video analysis. *Teaching Education*, *25*(4), 349-374.

Cochran-Smith, M., & Lytle, S. L. (2009). *Inquiry as stance: Practitioner research for the next generation*. New York, NY: Teachers College Press.

Cohen, D. (2004). *Teachers' professional development and the elementary mathematics classroom: Bringing understanding to light*. Mahwah, NJ: Erlbaum.

Coles, A. (2014). Mathematics teachers learning with video: The role, for the didactician, of heightened listening. *ZDM*, *46*(2), 267-278.

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

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

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, Institute of Education Sciences, U.S. Department of Education.

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

Garet, M. S., Wayne, A., Stancavage, F., Taylor, J., Walters, K., Song, M., . . . Doolittle, F. (2010). *Middle School Mathematics Professional Development Impact Study: Findings after the first year of implementation (NCEE 2010-4009)*. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Gaudin, C., & Chaliès, S. (2015). Video viewing in teacher education and professional development: A literature review. *Educational Research Review*, *16*, 41-67.

Goldsmith, L. T., & Seago, N. (2011). Using classroom artifacts to focus noticing. In M. G. Sherin, V. R. Jacobs, & R. A. Philipp (Eds.), *Mathematics teacher noticing: Seeing through teachers' eyes* (pp. 169-187). New York, NY: Routledge.

Hargreaves, A. (1996). Transforming knowledge: Blurring the boundaries between research, policy, and practice. *Educational Evaluation and Policy Analysis*, *18*(2), 105-122.

Harlin, E.-M. (2014). Watching oneself teach—Long-term effects of teachers' reflections on their video-recorded teaching. *Technology, Pedagogy, and Education*, *23*(4), 507-521.

Heller, J. I., Daehler, K. R., Wong, N., Shinohara, M., & Miratrix, L. W. (2012). Differential effects of three professional development models on teacher knowledge and student achievement in elementary science. *Journal of Research in Science Teaching*, *49*(3), 333-362.

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

Hill, H. C., Beisiegel, M., & Jacob, R. (2013). Professional development research: Consensus, crossroads, and challenges. *Educational Researcher*, *42*(9), 476-487.

Hill, H. C., Blunk, M., Charalambous, C., Lewis, J., Phelps, G. C., Sleep, L., & Ball, D. L. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction*, *26*, 430-511.

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

Hixon, E., & So, H. J. (2009). Technology's role in field experiences for preservice teacher training. *Journal of Educational Technology & Society*, *12*(4), 294-304.

Kennedy, M. M. (1999). Form and substance in mathematics and science professional development. *NISE Brief*, *3*(2), 1-8.

Kersting, N. (2008). Using video clips of mathematics instruction as item prompts to measure teachers' knowledge of teaching mathematics. *Educational and Psychological Measurement*, *68*, 845-861. doi:10.1177/0013164407313369

Kersting, N. B., Givvin, K. B., Sotelo, F. L., & Stigler, J. W. (2010). Teachers' analyses of classroom video predict student learning: Further explorations of a novel measure of teacher knowledge. *Journal of Teacher Education*, *61*(1-2), 172-181.

Kersting, N. B., Givvin, K. B., Thompson, B. J., Santagata, R., & Stigler, J. W. (2012). Measuring usable knowledge: Teachers' analyses of mathematics classroom videos predict teaching quality and student learning. *American Educational Research Journal*, *49*(3), 568-589.

Kiemer, K., Gröschner, A., Pehmer, A.-K., & Seidel, T. (2014). Teacher learning and student outcomes in the context of classroom discourse: Findings from a video-based teacher professional development program. *Form@re*, *14*(2), 51-62.

Kleinknecht, M., & Schneider, J. (2013). What do teachers think and feel when analyzing videos of themselves and other teachers teaching? *Teaching and Teacher Education*, *33*, 13-23.

Loucks-Horsley, S. (1996). Principles of effective professional development for mathematics and science education: A synthesis of standards. *NISE Brief*, *1*(1), 1-6.

MacDonald, E. (2011). When nice won't suffice: Honest discourse is key to shifting school culture. *Journal of Staff Development*, *32*(3), 45-47, 51.

MacDonald, E., & Shirley, D. (2009). *The mindful teacher*. New York, NY: Teachers College Press.

Mason, J. (2011). Noticing: Roots and branches. In B. Jaworski & T. Wood (Eds.), *Handbook of mathematics teacher education: Vol. 4. The mathematics teacher educator as a developing professional* (pp. 35-50). Rotterdam, The Netherlands: Sense Publishers.

Mitchell, R. N., & Ariemma-Marin, K. (2015). Examining the use of a structured analysis framework to support prospective teacher noticing. *Journal of Mathematics Teacher Education*, *18*(6), 551-575.

Montgomery, D. C. (1991). *Design and analysis of experiments*. New York, NY: Wiley.

Richardson, V. (2003). The dilemmas of professional development. *Phi Delta Kappan*, *84*(5), 401-406.

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. J., Garnier, H. E., Chen, C., Lemmens, M., Schwille, K., & Wickler, N. I. (2011). Video-based lesson analysis: Effective science PD for teacher and student learning. *Journal of Research in Science Teaching*, *48*(2), 117-148.

Roth-McDuffie, A., Foote, M. Q., Bolson, C., Turner, E. E., Aguirre, J. M., Bartell, T. G., . . . Land, T. (2013). Using video analysis to support prospective K-8 teachers' noticing of students' multiple mathematical knowledge bases. *Journal of Mathematics Teacher Education*, *17*(3), 245-270.

Sample-McMeeking, L., Orsi, R., & Cobb, R. B. (2012). Effects of a teacher professional development program on the mathematics achievement of middle school students. *Journal for Research in Mathematics Education*, *43*(2), 159-181.

Santagata, R., Kersting, N., Givvin, K. B., & Stigler, J. W. (2011). 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. A. (1983). *The reflective practitioner: How professionals think in action*. London, England: Temple Smith.

Seago, N. M., Jacobs, J. K., Heck, D. J., Nelson, C. L., & Malzahn, K. A. (2014). Impacting teachers' understanding of geometric similarity: Results from field testing the Learning and Teaching Geometry professional development materials. *Professional Development in Education*, *40*(4), 627-653.

Seidel, T., Stürmer, K., Blomberg, G., Kobarg, M., & Schwindt, K. (2011). Teacher learning from analysis of videotaped classroom situations: Does it make a difference whether teachers observe their own teaching or that of others? *Teaching and Teacher Education*, *27*(2), 259-267.

Shavelson, R. J. (1996). *Statistical reasoning for the behavioral sciences*. Boston, MA: Allyn and Bacon.

Sherin, M. G. (2003). Using video clubs to support conversations among teachers and researchers. *Action in Teacher Education*, *24*(4), 33-45.

Sherin, M. G., & Han, S. Y. (2004). Teacher learning in the context of a video club. *Teaching and Teacher Education*, *20*(2), 163-183.

Sherin, M., & van Es, E. (2005). Using video to support teachers' ability to notice classroom interactions. *Journal of Technology and Teacher Education*, *13*(3), 475-491.

Snoeyink, R. (2010). Using video self-analysis to improve the "withitness" of student teachers. *Journal of Digital Learning in Teacher Education*, *26*(3), 101-110.

Star, J., & Strickland, S. (2008). Learning to observe: Using video to improve preservice mathematics teachers' ability to notice. *Journal of Mathematics Teacher Education*, *11*(2), 107-125.

Stoll, L., Bolam, R., McMahon, A., Wallace, M., & Thomas, S. (2006). Professional learning communities: A review of the literature. *Journal of Educational Change*, *7*(4), 221-258.

Sun, J., & van Es, E. A. (2015). An exploratory study of the influence that analyzing teaching has on pre-service teachers' classroom practice. *Journal of Teacher Education*, *66*(3), 201-214.

Tripp, T. R., & Rich, P. J. (2012). The influence of video analysis on the process of teacher change. *Teaching and Teacher Education*, *28*(5), 728-739.

van Es, E. A. (2012). Examining the development of a teacher learning community: The case of a video club. *Teaching and Teacher Education*, *28*(2), 182-192.

van Es, E. A., & Sherin, M. G. (2006). How different video club designs support teachers in "learning to notice." *Journal of Computing in Teacher Education*, *22*(4), 125-135.

van Es, E. A., & Sherin, M. G. (2010). The influence of video clubs on teachers' thinking and practices. *Journal of Mathematics Teacher Education*, *13*(2), 155-176.

van Es, E. A., Tunney, J., Goldsmith, L. T., & Seago, N. (2014). A framework for the facilitation of teachers' analysis of video. *Journal of Teacher Education*, *65*(4), 340-356.

Walkoe, J. (2015). Exploring teacher noticing of student algebraic thinking in a video club. *Journal of Mathematics Teacher Education*, *18*(6), 523-550.

Wayne, A. J., Yoon, K. S., Zhu, P., Cronen, S., & Garet, M. (2008). Experimenting with teacher professional development: Motives and methods. *Educational Researcher*, *37*(8), 469-479.

Zhang, M., Koehler, M. J., & Lundeberg, M. (2015). Affordances and challenges of different types of videos for teachers' professional development. In B. Calandra & P. Rich (Eds.), *Digital video for teacher education* (pp. 147-163). New York, NY: Routledge.

Zhang, M., Lundeberg, M., Koehler, M. J., & Eberhardt, J. (2011). Understanding affordances and challenges of three types of video for teacher professional development. *Teaching and Teacher Education*, *27*(2), 454-462.

Author Biographies

Mary Beisiegel, PhD, is an assistant professor in the Department of Mathematics at Oregon State University. Her research interests include the use of video recordings of mathematics instruction in the context of professional development and studying the quality of mathematics in classrooms.

Rebecca Mitchell, EdD, is an assistant professor of mathematics at Pine Manor College, Chestnut Hill, MA. Her scholar interests include supporting teachers for instructional improvement and improving undergraduate mathematics access and achievement for students of color.

Heather C. Hill is the Jerome T. Murphy professor in education at the Harvard Graduate School of Education. Her primary work focuses on teacher and teaching quality and the effects of policies and programs aimed at improving both.