Using Tablets to Explore the Potential for Video-based Classroom Observations for Research and Professional Development

David Blazar[1], Barbara Gilbert[2], Corinne Herlihy[2], Claire Gogolen[2]
[1]University of Maryland
[2]Harvard University, CEPR

To cite this article:

Using Tablets to Explore the Potential for Video-based Classroom Observations for Research and Professional Development

David Blazar, Barbara Gilbert, Corinne Herlihy, Claire Gogolen

Article Info	Abstract
	Easy access to low-cost, high definition video technology has transformed how researchers and practitioners are approaching school improvement efforts. Video-based observations of classroom instruction provide researchers the means to define and measure aspects of effective teaching and suggest subsequent professional learning. In this article we describe three use cases for video observations based on the work we are doing at the Center for Education Policy Research (CEPR) at Harvard University. The cases discuss strategies for (1) using teacher videos to improve instructional practice through 1:1 coaching focused on the use of a mathematics-specific observation rubric, (2) making observation and feedback cycles practical for over-burdened principals, and (3) developing outcome measures to inform our understanding of teacher effectiveness and suggest the potential for improvement. We describe a research agenda that is grounded in concrete examples of teaching practice.

Introduction

For more than two decades, U.S. educational policy and research has been directed at improving student achievement. Increased access to school-level data during this time, has focused attention on understanding how teachers impact student test score performance, and, subsequently, defining and measuring aspects of quality instruction (Jackson, Rockoff, & Staiger, 2014). A key objective for this work has been the potential to provide teacher feedback and suggest strategies for instructional improvement (Goe, Bell & Little, 2008) that might lead to increased teacher effectiveness and ultimately increased student learning.

More recently, growing availability to emerging video-based technologies such as iPads and Android tablets has changed how researchers and practitioners approach this work. Those interested in instructional improvement have begun exploring the use of video as a tool to inform teaching. Easy access to inexpensive, high-quality video cameras has made it possible to capture varied examples of classroom teaching over time. An affordance of video-based observations is that they serve to anchor notions of teacher effectiveness by grounding discussions in concrete examples of teaching practice. Video is asynchronous and thus, it allows researchers (and practitioners) an opportunity to pause when observing complex teaching environments, providing time to fine-tune our thinking and expressions of effective teaching. As a result, researchers are using video-based observations to define key characteristics of instruction, support instrument development and measure effective teaching in an effort to understand the relationship between teacher practice and student outcomes and then, to suggest improvements.

We see video as a helpful – likely essential tool – that can be used to improve teaching at several levels of organizational change, including classrooms, schools, and educational agencies. When considering classroom practice, video provides a detailed, objective record of what occurs in teachers' classrooms that can help teachers and observers identify areas for growth and potential for coaching. In contrast to in-person observations, video observations help make observation and feedback cycles sustainable for over-burdened school administrators. When used as an outcome measure, video is arguably the only way to identify differences across educational systems (e.g., districts, states, and countries), as other measures of teacher quality, such as value-added, are normative and rank teachers within a single given context.

Method

For this article, we present three use cases for video, based on our work at the Center for Education Policy Research (CEPR) at Harvard University. The cases provide an overview for how our work is evolving around the use of video. Due to space limitations, we are unable to provide detailed information about the research designs for each study mentioned. However, we do point to working papers and published resources that discuss the methodology and subsequent findings in detail and readers can learn about our work and access additional resources on the CEPR website: https://cepr.harvard.edu

Overview of the Selected Cases

For the first use case, we describe how we are using teachers' own videos to inform and guide professional learning through 1:1 coaching. The idea for the coaching work came from a pilot study that investigated the conditions needed for effective video-based professional development and the potential for using a content-focused observation instrument to provide a common language for teachers to discuss their practice. During the pilot, elementary mathematics teachers (grades 3-8) received summer training on the Mathematical Quality of Instruction (MQI) observation rubric and participated in ten afterschool discussion sessions focused on the MQI and videos of mathematics classroom practice. Early findings from this work were promising and led to the "Developing Common Core Classrooms Through Rubric-Based Coaching" study funded through the National Science Foundation. This later work included development and implementation of what came to be known as the "MQI Coaching Cycle" and included a rigorous evaluation of the program's effectiveness through application of a random assignment design at the teacher level. One hundred and forty-two mathematics teachers (grades 3-8) were recruited from two Midwestern public school districts. Seventy-two teachers were assigned to an MQI Coaching (treatment group) and 70 teachers were assigned to the control group. The majority of participating teachers completed 10 or more coaching cycles. Early findings for the study can be found at http://cepr.harvard.edu/files/cepr/files/mqi-coaching-research-findings.pdf.

The second use case describes how we have used video to make observation and feedback cycles viable for principals. The Best Foot Forward (BFF) project investigated whether video-based observations would make the classroom evaluation process easier to implement, less costly, and more valid and reliable. In a randomized controlled trial, the study team provided each teacher an iPad and a Swivl robot and asked them to videorecord examples of regular classroom practice. Teachers were allowed to select their "best" lessons for evaluation. At question was whether digital video made the observation process more acceptable to teachers and administrators. In 2013, the project was piloted in 100 classrooms in New York City, Georgia, and North Carolina. More than 400 teachers and their administrators from districts in Delaware, California, Colorado, and Georgia joined the impact evaluation from 2013–15. The year one implementation brief provides additional information and is available here,
 https://cepr.harvard.edu/files/cepr/files/l4a_best_foot_forward_research_brief1.pdf.

And finally, the third case discusses how video is being used as outcome measures to benchmark the characteristics of effective teaching across varied contexts, with the goal of informing school improvement efforts. We begin this section with a discussion of two early examples from the field of video-based observations as measures of effective teaching, including the Third International Mathematics and Science Study (Stigler et al., 1999) and the Measures of Effective Teaching Study (Kane, McCaffrey, Miller, & Staiger, 2013). We then share three examples of studies conducted at CEPR beginning with the Mathematics Teachers and Teaching Survey (MTTS). Modeled after the TIMSS, the MTTS study took a "high-touch" approach to gathering video-based observations from 200 randomly selected U.S. middle school mathematics teachers. Videos were coded using the Mathematical Quality of Instruction observation instrument (see the section below for a brief discussion of the observation instrument) Next, we discuss the Developing Measures of Mathematics Teaching, the core study from the IES-funded National Center for Effective Teaching that gathered four measures of effective teaching from 300 randomly selected teachers from four U.S. school districts over three years. And finally, we discuss an evaluation study of the Math Solutions video-based professional development program as it was implemented in one urban school district. Again, a thorough discussion of the methodology used in each of these three studies is not possible within the constraints of this article, however, we do provide links to working papers and study findings.

Use of the Mathematical Quality of Instruction Observation Instrument

Like other measurement tools, video – and the instruments and protocols used to evaluate it – must have strong measurement properties in order to be able to be used to draw valid inferences regarding instruction. In two of three cases described below, we rely on the Mathematical Quality of Instruction (MQI) observation instrument, which was developed in part by the research teams who also contributed to the cases described below (See https://cepr.harvard.edu/mqi-research-basis). The teaching skills and practices captured by the MQI align closely with decades worth of theory and research on "ambitious" and cognitively demanding instruction in mathematics (Hill et al., 2008; National Council of Teachers of Mathematics [NCTM], 1989, 1991, 2000). Elements of instruction – including the extent to which teachers link across multiple representations of mathematical ideas and provide opportunities for students to solve complex problems such as looking for patterns – also are similar to those identified by the recent Common Core State Standards (National Governors Association for Best Practices, 2010). Research projects using the MQI instrument indicate that it adequately captures the quality of teachers' instruction when used by multiple raters to score multiple lessons (Hill, Charalambous, & Kraft, 2012). Scores generated from the instrument have been shown to relate to other measures of teacher quality including math content knowledge and mathematical knowledge for teaching (Hill, Blazar, & Lynch, 2015; Hill et al., 2008), as well as student outcomes including test score performance (Blazar, 2015b) and students' self-efficacy in math (Blazar, & Kraft, 2017). These properties are critical for using a tool such as the MQI to draw valid inferences regarding instruction.

Several other research projects, including our second use case, have successfully paired video technology with additional observation instruments and protocols (e.g., Framework for Teaching, Classroom Assessment Scoring System [CLASS], Protocol for Language Arts Teaching Observation [PLATO]). We encourage readers to examine other cases beyond the three described below and to consider how video technology and appropriate protocols and observation instruments can be used as a means to improve instruction and student outcomes, including in other content areas beyond mathematics.

Case 1: Video-based Professional Learning

Scope and Theory

Much work has been done over the last two decades to improve U.S. mathematics instruction. Reform standards for instructional practice such as those defined by NCTM principles for mathematics instruction and the Common Core State Standards now require teachers to teach conceptually and implement new teaching methods thought to be advance student learning. However, implementing these reforms with fidelity is dependent on teachers' knowledge of content and pedagogy (Carpenter, Fennema, Peterson, Chiang, & Loef, 1989; Fennema, Franke, Carpenter, & Carey, 1993; Kazemi & Franke, 2004; Lamon, 2012) and corresponding ability to translate reform standards to instructional practice (Ball & Cohen, 1999; Stigler & Hiebert, 1997). Teachers also must have sufficient mathematical domain knowledge to interpret formative data gathered during student-teacher and student-student interactions and adapt instruction to address individual student needs (D.L. Ball, Thames, & Phelps, 2008; Black & Wiliam, 1998, 2009).

Although schools and districts dedicate time and money for teacher professional development (PD), studies of PD interventions do not show consistently positive results (e.g. Garet et al., 2008; Jacob, Hill & Cory, 2017; Yoon et al., 2007). However, one bright spot in the literature is focused on instructional coaching. Several recent studies demonstrate coaching models to be effective at improving instructional quality and subsequent student outcomes (Allen, Pianta, Gregory, Mikami, & Lun, 2011; Campbell & Malkus, 2011; Kraft & Blazar, 2014; Taylor & Tyler, 2012).

Instructional coaching can refer to a wide variety of interventions and supports for teachers, but a review of the existing literature suggests several key features may be needed for coaching to be effective. First, the focus of coaching must be on *classroom observation and feedback over time* and coaching discussions should be focused on how to improve instruction. Second, effective instructional coaching features *high quality interactions between teachers and coaches*. Finally, the *coaches must be highly trained*. Campbell and Malkus (2011) found that highly trained mathematics specialists, receiving more than a year of training before beginning their work coaching grades 3-5 teachers, made a significant, positive difference in student achievement over time. This approach contrasts to that taken by school districts that sometimes promote a skillful classroom teacher to the role of instructional coach without specialized training in mathematics content or pedagogy. The authors further

note that the results of coaching on student achievement cannot be generalized to contexts that do not provide substantive training.

Using Video to Improve Instructional Practice

The first use case describes our use of classroom video as a tool for teachers' professional growth employing a 1:1 coaching model. MQI Coaching – a mathematics-specific, video-based coaching model – helps teachers make incremental changes in their instruction, with the long-term goal of improving mathematics instruction in a realistic and sustainable way (See https://mqicoaching.cepr.harvard.edu/coaching-cycle). MQI Coaching relies heavily on the use of video and the MQI observation rubric to achieve these criteria. Groups of teachers in Grades 3–8 received summer training on the MQI rubric and participated in ten afterschool discussion sessions based on the MQI and video of math instruction. Teacher groups were randomly assigned to a type of facilitation (high-facilitation or teacher-led) and type of video use (stock video or video of their own instruction).

Classroom video drives the success of the MQI Coaching model in several ways. In each two-week cycle, teachers and coaches watch and analyze three short video clips of instruction: two of the teacher's own instruction and one example clip from the MQI video library. This video review provides teachers with an exposure to a wide range of practice, as well as a substantial amount of practice using the MQI rubric to unpack the enacted lesson as it unfolds in the classroom context. We have developed a structure and set of routines for coaching that allow us to leverage both video and the MQI rubric to produce high-quality conversations about instruction and improvement. MQI coaches require advanced mathematical content knowledge to successfully complete training on the content-focused rubric. After the initial training, all coaches receive continuous training and support focused on their delivery of the MQI routines and ability to facilitate successful, targeted conversations about instruction with teachers.
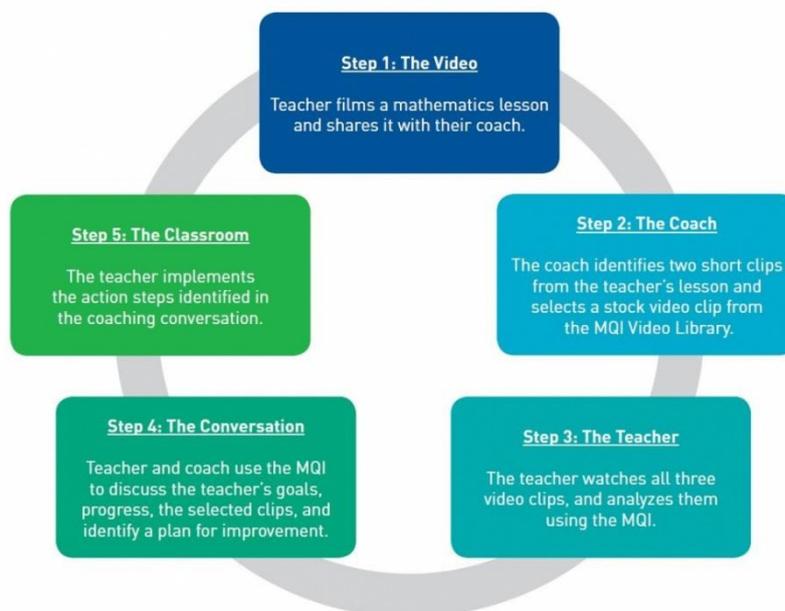
Figure 1: MQI Coaching Cycle

For each video clip, coaches facilitate a discussion in which teachers first ***describe*** and then **elevate** the instruction in the clip, staying focused on just one or two MQI codes at a time. After the teacher and coach describe and elevate all three video clips, the teacher selects one or two realistic and actionable *next steps* to try out. The conversation structure consisting of *describe*, then *elevate*, then select *next steps* provides structure and routines that support coaches and teachers in having high-quality coaching conversations that remain focused on instruction. Using the MQI to analyze short video clips, coaching conversations remain focused and specific; teachers walk away with increased understanding of what occurred during instruction and what could be improved.

This process for focused and specific analysis of instruction is built into the regular routine of MQI Coaching conversations, keeping the coaching conversations focused on improvements to instruction and helping keep the coaching conversations high-quality. It is important to note that facilitating such high-quality conversations is not an easy task. Just as we would not expect a classroom teacher to facilitate a high-quality mathematics lesson without significant training, planning, and preparation, we do not expect that coaches will be able to facilitate high-quality coaching conversations without substantial training and planning.

To this end, we provide coaches with substantial training and ongoing support. Our initial coach training is intensive, with coaches receiving training on relationship building and practice in facilitating MQI-grounded conversations that progress through the "describe and then elevate" portions of the coaching conversation. To support the "lesson planning" conversation, we provide coaches with a conversation planning guide, which provides guiding questions coaches can use to help them select video clips and prepare the questions they will ask teachers. In addition to an intensive initial coach training, MQI coaches participate in ongoing coach support sessions, in which coaches discuss common challenges and describe strategies for success. Coaching conversations are also recorded, so our coach support protocol includes ongoing listening-in oversight, in order to identify and target areas for coach growth. By providing intensive and ongoing training and support for coaches, we can offer higher-quality coaching conversations to all participating teachers.

Piloting MQI Coaching

We rolled out the 1:1 coaching model as a pilot study in two mid-western school districts. We hired certified MQI raters to be our coaches, and provided extensive training on how to use the MQI as a coaching tool. Trained coaches met with participating teachers regularly. Coaching conversations were tightly focused on the MQI rubric, analysis of the video and reflecting on mathematics instruction. During each session, teacher and coach discussed one stock video from our video library, and two video clips of the teacher's own classroom. The coaches received ongoing oversight and PD on how to maintain high quality conversations. Doing so, created the conditions that the literature suggests are important for instructional coaching to work.

As teachers progressed through this intensive coaching program, their approach to classroom instruction gradually began to incorporate a growing understanding of the MQI dimensions. Teacher discourse revealed how they were beginning to think about the MQI when planning lessons. Teachers discussed instructional decisions such as how to respond to students and when to intervene. For example, one participant described her learning experience,

> It has not always been easy to watch video of myself and analyze my teaching, but there is no question that it has made a dramatic impact on my teaching and the way I approach teaching. It's incredible. I was just telling my coach that I am thinking about the MQI during my lesson, not just as I reflect afterwards. I'm thinking to myself things like, "This would be a good opportunity to have the students explain something" or "Okay. I have this student is giving an explanation, but can I probe deeper." Also, I'm seeing concrete results from the students having that deeper understanding of math concepts.

In the randomized trial of 1:1 MQI Coaching, teachers who received coaching reported more instructional improvement than teachers in the control group who did not receive the coaching (see https://cepr.harvard.edu/files/cepr/files/mqi-coaching-research-findings.pdf for early findings). Importantly, this self-report data agrees with findings from video of their instruction. We returned one year after the coaching intervention to videorecord and score an additional five lessons each from a random selection of teachers in both the treatment and control groups. Teachers in the treatment group scored significantly higher than control group teachers on three MQI domains, including Common Core-Aligned Student Practices, Working with Students and Mathematics, and Richness of the Mathematics.

Research on instructional coaching (Kraft, Blazar, & Hogan, 2017) suggests it may be most effective when several criteria are met: coaches are highly trained, coaching is focused on observation and feedback over time, and the conversations between teachers and coaches are high-quality (For the updated working paper, see https://scholar.harvard.edu/files/mkraft/files/kraft_blazar_hogan_2016_teacher_coaching_meta-analysis_wp_w_appendix.pdf). MQI Coaching uses video to achieve all three of these aims. Results from the randomized trial suggest that MQI Coaching does result in improved instruction. Video supports this work in several ways. Video recording the lesson allows the coach and teacher to watch and re-watch the video and analyze the instruction, rather than relying on often unreliable memories. Using the MQI rubric and evidence

from the video ensures coaching conversations remain focused, specific, and high-quality. Ongoing coach training uses classroom video examples and recordings of coaching conversations to help coaches improve as facilitators.

Case 2: Video Technology to Improve School-Based Teacher Evaluation

Scope and Theory

A second way that video-based technology is essential to efforts to improve teaching quality is by making observation and feedback cycles sustainable at the school level. For decades, principals and school administrators have been called on to evaluate teachers using on-the-job performance measures in an effort to improve instruction and, in turn, student outcomes (see, for example, Darling-Hammond, Wise, & Pease, 1983; Gallup, 1979; Gudridge, 1980). As described earlier in this paper, observations of teaching practice with standards-based rubrics provide an opportunity to identify areas for improvement and to match teachers to professional development (Danielson & McGreal, 2000; Darling-Hammond, 2013; Hill & Grossman, 2013; Odden, 2004; Papay, 2012; Wright, Horn, & Sanders, 1997). Formal evaluations that incorporate observations of teaching practice may also be used to identify teachers for dismissal, which might improve the quality of schooling if these ineffective teachers are replaced with more effective ones (Hanushek, 2009). Following efforts under the Obama administration to make teacher evaluation a cornerstone of educational reforms, more than forty states have enacted new legislation aimed at strengthening and expanding teacher evaluation systems in public schools (National Council on Teacher Quality, 2013). In particular, schools have been asked to increase the amount of feedback teachers received and use these data to inform decisions on professional development, promotion, tenure, and dismissal (Hallgren, James-Burdumy, & Perez-Johnson, 2014; Partee, 2012).

Despite calls such as these, however, teacher evaluation historically has not achieved its stated mission. In 2009, The New Teacher Project's report, "The Widget Effect," documented widespread failure of evaluation systems to differentiate teachers with regard to performance standards, let alone to act upon them in making job decisions. Across twelve school districts, less than 1% of teachers received an unsatisfactory rating, and over 94% earned top ratings. Only 26% of teachers received areas for instructional development as part of their evaluations. Less than 0.1% of tenured teachers were dismissed due to performance. These glaring statistics led the authors to conclude that, "school districts fail to acknowledge or act on differences in teacher performance almost entirely. When it comes to officially appraising performance and supporting improvement, a culture of indifference about the quality of instruction in each classroom dominates" (Weisberg, Sexton, Mulhern, & Keeling, 2009, p. 2). These patterns live on today, even after several years of implementation of the Obama-era teacher evaluation reforms (Kraft & Gilmour, 2017).

The failure of teacher evaluation is due to several key constraints placed on schools and their leaders. First, leaders must be able to observe instruction in a way that leads to reliable inferences about teachers' practice. At a minimum, this requires that principals have access to and ability to use measurement tools (e.g., observation instruments) that capture the most important features of instruction and to differentiate teachers with regard to their ability to implement these practices (Bell et al., 2012; Hill et al., 2012; Kane & Staiger, 2012). Even if they have access to these tools, though, there is concern that principals and school leaders may have limited expertise in unique content areas such as mathematics (Hill & Grossman, 2013), which may undermine the observation process if teachers' perceive the scores and feedback they receive to be unreliable (Herman & Baker, 2009). Principals' discomfort in giving low ratings, even if they perceive teachers to perform below expectations (Kraft & Gilmour, 2016), may be related to their discomfort in justifying these scores to teachers.

Further, schools must have the capacity to successfully implement observation-based evaluation systems in light of considerable financial and human capital investments (Bridges, 1990; Chambers, Brodziak de los Reyes, & O'Neil, 2013; Range et al., 2011). One of the key investments, principals' time, has proven to be a considerable barrier. Two recent studies suggest that principals spend only about 10% of their time on instructional activities, and even less time (roughly 2%) on formal evaluations of teaching (Grissom, Loeb, & Master, 2013; May & Supovitz, 2011). Most frequently, principals engaged in brief classroom walkthroughs, rather than lengthier observations of teaching recommended by researchers and policymakers. Coaching activities, in which principals used observations of classroom practice to identify ways for teachers to improve their practice, were even more rare.

Substituting Video-based Evaluation for In-person Observation

As described elsewhere (Kane, Gehlbach, Greenberg, Quinn, & Thal, 2015), we evaluated one strategy thought to address these key challenges to the teacher evaluation model: allowing teachers to submit their own recorded lesson videos in lieu of in-person observations (Jacobs, Doherty, Lakis, Lasser, & Staresina, 2014). We hypothesized that digital video would offer a number of potential advantages over in-person observations. Video would provide a more detailed, objective record of what occurred in teachers' classrooms than an observer's written notes. In turn, feedback cycles could focus on this record of teachers' classroom practice in order to identify areas of strength versus areas of weakness in need of improvement. Reliance on objective evidence could also help build trust between teachers and principals. For example, teachers may be more assured that evaluators' feedback reflected what actually happened in the classroom. Finally, video would allow principals to time-shift their observational duties to quieter times of the day or week, and facilitate the use of external observers and content experts. Ultimately, we hypothesized that evaluation feedback would be most effective if paired with a video record. (See the full report at https://cepr.harvard.edu/publications/best-foot-forward-project-substituting-teacher-collected-video-person-classroom .)

We tested this hypothesis through an experimental evaluation with a sample of teachers (N = 433) across four U.S. states. Treatment teachers were given access to a video camera and a secure website for uploading, viewing, sorting, and sharing these recorded lessons. In the first of two cohorts, teachers received camera kits and technical support from the hardware supplier, thereNow. The cameras incorporated two video streams (one for the teacher and one for students) and three audio channels (one for the teacher and two for general classroom audio). At the end of each lesson, the portable device merged the video and audio streams into a single video file. In cohort two, teachers used a recording device from Swivl, which attached to an iPad mini. In both cohorts, a private contractor, BloomBoard, provided video storage and the platform to navigate these videos.

In the secure platform, teachers could rewatch all of their recorded lessons and then identify those they wished to share with their principals or school leaders for the purpose of evaluation. Administrators did not have access to any videos other than those that the teacher shared with them. After a teacher shared a video, the administrator logged in, tagged specific moments of the video, and provided written comments to accompany these tags and other key moments in the lesson. The software was customized so that the tags would correspond to each district's formal observation rubric, which were aligned closely to the Danielson Framework. Key dimensions of these instruments included the extent to which teachers planned and prepared appropriately for instruction, the quality of their instructional delivery, and the quality of the classroom environment. During playback, the observer's comments would appear at the specific point in the video when the observer entered them. The observer then shared the video evidence and commentary with the teacher before the two parties sat down in person to discuss the video feedback and determine a final score.

Comparing the efficacy of this approach to in-person classroom observation and feedback cycles, we found that video-based evaluation resulted in more productive feedback, more collegial relationships, and stronger working environments. Directly addressing concerns around time commitments related to teacher evaluation, this approach led to an increase of 4.5 minutes per week of observation than the control group, as reported by principals, as well as a decrease in other administrative aspects of the observation process including completing forms. As reported on an end-of-year survey, teachers in the treatment group perceived their supervisors to be more supportive and their observations to be fairer than those in the control group. For example, treatment teachers were statistically significantly less likely to report that their conversations had been adversarial or that they disagreed with the administrator about the appropriate score. Principals generally agreed with this assessment. For example, administrators in the treatment group were statistically significantly more likely to report that teachers were "never" or "rarely" defensive during the post-observation conference. (For a complete description of the results of the study, see Kane et al., 2015.)

The design of the evaluation meant that it was not possible to compare the instructional quality of treatment versus control group teachers, as occurs in other cases described in this paper. This is because only treatment teachers recorded lessons that could then be scored by outside observers blind to treatment status; recording lessons for control-group teachers would have contaminated the treatment-control contrast we sought to establish in the experimental design. At the same time, self-reports by teachers indicated that treatment teachers were more likely to describe a specific change in their practice resulting from their post-observation conference, including their work around time management and lesson pacing. Interestingly, the intervention also resulted in teachers becoming more self-critical of their teaching. Treatment teachers rated their instructional practice related to classroom management, engagement of students, and other areas as statistically significantly lower than control group teachers. These are all skills that would be observable in a video recording.

In addition to resulting in changes in self-reported behaviors, the intervention also generated positive effects on some observed behaviors. At the end of the treatment year, treatment teachers were more likely to remain in their same teaching assignment – in their same school, grade, or district. This type of stability in teachers' assignments has been found to be positively related to their growth in effectiveness over time and to improved student outcomes (Atteberry et al., 2017; Blazar, 2015a; Ost, 2014). Despite positive effects on teachers' perceptions of the evaluation process and of their behavior, the intervention did not appear to improve students' academic performance as we originally hypothesized it might. However, we see these results as consistent with several other evaluations of teacher-level interventions, such as teacher professional development, that found changes in proximal outcomes (i.e., teacher knowledge or behavior) but not more distal outcomes (e.g., student performance on standardized assessments) (Garet et al., 2008; Garet et al., 2011; Garet et al., 2016; Glazerman et al., 2010).

Case 3: Video as an Outcome Measure

Scope and Theory

In this third and final use case, we focus on how video facilitates a stronger, more reliable metric of instruction that allows researchers to benchmark examples of effective teaching practice nationally and internationally and to track the results of improvement efforts over time. While past efforts of this sort have come primarily from researchers, we see several opportunities for education agencies to track progress as a complete system. Video-based teacher observations increasingly serve as outcome measures for the purpose of providing insights into complex classroom settings. Over the past decade, researchers have begun developing instruments to define benchmarks and measure the characteristics of effective teaching (e.g. Blazar, 2015b; Hill, 2007; Hill et al., 2012; Hill, Schilling, & Ball, 2004), rate instructional practice as a measure of teacher quality (e.g. Cantrell & Kane, 2013; Goe, Bell, & Little, 2008); and apply the findings to the development of interventions to improve teacher practice (e.g. Borko, Jacobs, Eiteljorg, & Pittman, 2008; Hiebert & Stigler, 2000; Santagata, 2009; Sherin & van Es, 2005). To highlight the complexity of this work, we begin this section with a brief historical review of how video has emerged as a viable research tool.

The 1995 TIMSS video study of mathematics and science teaching (Stigler et al., 1999) was a first example of how video might be used to conduct surveys of teaching practice for the purpose of comparison. In the initial study, TIMSS researchers examined videos of eighth grade mathematics instruction from classrooms in the U.S., Germany, and Japan to determine if instructional reforms were being implemented in classrooms. For the purpose of comparison, researchers identified individual classroom lessons as the unit of analysis and then, created a coding scheme to examine specific aspects of content and pedagogy. This novel use of video as an outcome measure clearly suggested a "gap" existed between teachers' professed understanding of reform practices and their enacted practice (Hiebert & Stigler, 2000), and also highlighted the potential for improvement. Since 1995, the TIMSS study has been conducted every four years and data collection has grown to include classroom video and student assessment scores for mathematics and science from more than sixty countries. The TIMSS study was groundbreaking for demonstrating that video could provide a richly descriptive view of instructional practice as well as a set of benchmarks for teacher quality.

Defining and measuring aspects of teacher effectiveness has continued to focus U.S. research efforts. Over the last ten years, researchers have worked to develop valid measures of teacher effectiveness and create systems of teacher evaluation that fairly differentiate teachers' skill and impact on student outcomes. This effort has been driven largely by the Obama administration's Race to the Top competitive grant program of state departments of education, which provided $4.35 billion to nineteen states that submitted plans to drastically revamp their teacher evaluation systems by including multiple measure of teacher effectiveness, include standardized measures of student achievement (U.S. DOE, RTTT). The attention to teacher effectiveness received strong support from the Bill and Melinda Gates Foundation (BMGF); BMGF funded the Measures of Effective Teaching (MET) Project (Kane, McCaffrey, Miller, & Staiger, 2013). The U.S. Department of Education's Institute of Education Sciences (IES) also turned significant funding to the research topic of teacher effectiveness with research grants and a national research and development center, the National Center for Teacher Effectiveness at Harvard University (See https://cepr.harvard.edu/ncte). These research projects relied significantly on video of classroom practice.

Using Video to Advance Our Understanding of Teacher Effectiveness and Improve Educational Systems

At the Center for Education Policy Research (CEPR) we have had several projects that rely on classroom observations captured on video as an outcome measure to (a) benchmark and measure the characteristics of effective teaching, particularly in mathematics; (b) rate instructional practice as a measure of teacher quality; and c) use these rating to evaluate the effectiveness of an intervention designed to support better teaching and improved student outcomes. The first example is the Middle School Mathematics Teachers and Teaching Survey (MTTS). The study aims to answer the question, "What is the current state of mathematics education in the United States?" Anecdotal evidence suggests that new teaching methods, technologies, and curriculum materials have appeared in U.S. mathematics classrooms over the last two decades. Many of these innovations are associated with principles for mathematics instruction developed by the National Council for Teachers of Mathematics (NCTM), as well as the Common Core State Standards. In 2015, CEPR began collecting data for the study designed to understand the consequences of these standards and curricular changes for middle school mathematics instruction. As part of this study, our team asked a randomly selected sample of 200 U.S. middle school mathematics teachers to self-videotape up to four lessons using a Samsung Galaxy 4 tablet and Swivl robot. Data collection included videos from 158 teachers from 43 states. The videos are analyzed using the MQI instrument. For each of the three domains measured, MQI raters assigned scores to 7.5-minute segments of instruction, with scores indicating that elements of the domain were not present, minimally present (low), present (mid), or present with extended/strong implementation (high).

A detailed description, along with early findings from the MTTS study (See https://cepr.harvard.edu/middle-school-mathematics-teachers-and-teaching-survey) will be of interest to mathematics teachers, teacher trainers, school districts and policymakers. Data collection was completed in August 2017 and we have begun final analysis; we anticipate the final report will be available in 2018. The mechanics of the study, particularly how video observation is translated to valid measures of teaching, is likely of most interest to the research community. It is important to note that this work was not exclusive to the MTTS study. In fact, most of the effort to validate the MQI instrument and devise a reliable scoring process occurred over ten years of prior studies. (For more information on the development and use of the MQI Instrument, please see https://cepr.harvard.edu/mqi-research-basis.)

Our second example focuses on the use of video to rate instructional practice as a measure of teacher quality. The National Center for Teacher Effectiveness (NCTE) included a core study, Developing Measures of Effective Mathematics Teaching. This study collected four measures of teacher effectiveness – classroom observations rated with the MQI; mathematics knowledge for teaching (teacher assessment); student engagement surveys; and value-added scores based on student achievement changes attributed to the teacher. These data were collected from nearly 300 teachers in 50 schools in four school districts over three school years.

The NCTE dataset, and the measure of classroom practice in particular, furthered our understanding of the relationship between measures of teaching effectiveness and the variation in teaching practice across settings. Because four school districts participated in the NCTE core study, we were also able to explore variation in the quality of teaching across settings (Hill, H.C, Blazar, D., and Lynch, K., 2015). The working paper, "What Does It Mean to be Ranked a 'High' or 'Low' Value-Added Teacher? Observing Differences in Instructional Quality Across Districts", can be found here: https://cepr.harvard.edu/publications/meaning-high-and-low-value-added-teaching-observing-differences-instructional. Our work also explored the comparability of test-based or value-added metrics across districts and the extent to which they capture variability in classroom practice (Blazar, D., Litke, E., & Barmore, J., 2016). We found positive though weak correlations between value-added and observation-based measures of teacher performance (Chin, M. & Goldhaber, D., 2015). The working paper, "Exploring Explanations for the 'Weak' Relationship Between Value Added and Observation-Based Measures of Teacher Performance", can be found here https://cepr.harvard.edu/publications/exploring-explanations-weak-relationship-between-value-added-and-observation-based. However, we were able to link specific instructional dimensions and changes in student achievement, finding that inquiry-oriented instruction positively predicts student achievement (Blazar, D., 2015).

This last finding is important to our final example of how video of classroom practice is used as an outcome measure. As noted, many studies have evaluated the effectiveness of interventions designed to support better teaching and improved student outcomes. Often the intervention's logic model envisions some change to curriculum and/or to the professional development provided teachers will lead to changes in teaching practice. These changes in practice will in turn lead to changes in outcome for students. One such study was conducted by researchers based at Harvard University and the University of Michigan. The three-year evaluation of the Math Solutions professional development program captured multiple outcome measures, including video of

classroom practice (See Jacob, R.T., Hill, H.C. & Corey, D., 2017). The study researchers and the program's developers mapped a professional development logic model that began with a needs assessment and professional development aligned to district needs. Next, the research team determined the content and dosage of professional development needed to lead to high quality mathematics instruction (e.g. engaging and high cognitive demand tasks, students developing their own solutions to mathematical problems) and ultimately, to improved student achievement. The research questions and data collection plan incorporated the following logic model: Was the Math Solutions professional development program implemented with fidelity? (observe PD sessions; track teacher participation); How did Math Solutions impact teachers' mathematical knowledge for teaching and the mathematical quality of their instruction? (MKT teacher assessment; video classroom observations); and How does Math Solutions professional development impact students' mathematical achievement? (student math assessments). The program was implemented with strong fidelity in the first year, but participation declined in the subsequent two years. The study found limited positive impact on teachers' mathematical knowledge for teaching, but no impact on the targeted instructional practice or student achievement. Across all three studies, we relied on a team of individuals or raters to score the classroom video. The raters completed a 20-hour training on the MQI instrument and demonstrated capacity to differentiate instruction along its four domains and more than 25 codes. An online training was developed (See https://cepr.harvard.edu/mqi) to prepare raters to score video using the MQI rubric.

Importantly, an observation instrument or rubric is not valid or reliable without a system of scoring that had been designed and tested to ensure fidelity of implementation. A series of generalizability studies guided the development of the scoring protocol and system (Hill, H.C., Charalambous, C.Y., and Kraft, M.A., 2012), that includes collecting at least three observations per year per teacher, scored by two raters, and with regular calibration of raters to assure reliability. This scoring system was implemented for MTTS, NCTE and the Math Solutions evaluation.

Conclusion

Increased access to high quality video is transforming the ways that we think about field-based research and the potential for professional learning in educational settings. Video has opened the doors to teachers' classrooms and possibility of viewing instructional practice in context. Doing so has afforded researchers the opportunity to study the characteristics of effective teaching and then, to apply that knowledge to professional learning. The three use cases presented above represent the ways in which we are using video-based observations to inform the development of outcome measures and as exemplars of instructional practice to support professional learning for principals and teachers.

As video technology has improved and become increasingly affordable, there is growing interest among school-based practitioners and researchers to think about how it might be used to inform our work. However, there are specific challenges that are important to address when considering the use of video. Below we suggest important considerations for the use of video and provide links to available resources.

Challenges and Recommendations

There are four challenges that are consistent across all three use cases presented earlier, including (a) cost of the equipment, (b) what can be observed, (c) availability of technical support, and (d) safety concerns. Below, we speak briefly to each of these concerns. For more detailed information we suggest reading the Best Foot Forward: Video Observation Toolkit (see https://cepr.harvard.edu/video-observation-toolkit). The toolkit will help you think about what is needed to successfully video record classroom practice in your context.

Cost of Equipment and Viewing Platform

The cost of quality camera equipment has decreased substantially in the last few years. A decade ago, high definition camera units suitable for research were expensive and unwieldy. Cameras are much less expensive than before; we currently budget between $800-1100 for each camera setup. Before purchasing a camera, you should consider the purpose of the video and how many units you will need. Think about what you want to capture (i.e. the board, the teacher, students talking, etc.) and what camera features are needed. For example, basic features that we look for when choosing a camera setup include a high definition video camera (or tablet), a high-quality microphone, and tripod. We increasingly use tablets as video cameras because they have the

capacity to capture high definition video even in low light settings. Another feature to consider is the audio quality and ability to hear student voices. We use an external wireless microphone in addition to the built in microphone on the tablet and sometimes add a second wireless microphone that can be placed in the center of the classroom or in student groups to capture student conversations.

Deciding where you will upload the video for viewing is also a consideration. There are a number of commercial viewing platforms currently available that provide features like the ability to timestamp and annotate the video. Alternatively, for in-person viewing a low-tech solution might be as simple as playing the video on the tablet. We encourage you to use the BFF toolkit to help you think through these types of decisions.

The Need for Technical Support

Choosing the right video camera equipment is important to ensure teacher participation. Equipment that is difficult to set up and use can limit teacher participation. It is also necessary to provide technical support to help set up the camera for novice users and troubleshoot malfunctioning equipment. We often elicit the help of the district IT staff to support the hardware and network infrastructure. When appropriate, we sometimes train a site coordinator on the use of the camera to provide support where needed. In addition, we provide appropriate "quick guides" that support camera setup and use, as well as a series of short, online videos that walk teachers and support staff through the camera setup and protocols for video recording. (See http://www.mqiprojects.org for examples of support material, including the video guides from the MTTS and MQI Coaching studies.)

Recruitment and Consent Process

When suggesting that teachers video record their practice, it is essential to articulate clearly how the videos will be used (i.e. for instructional learning, evaluation, etc.) Recognize that some teachers may find video recording to be somewhat intrusive and/or uncomfortable. Take care to describe the purpose and uses for the video to alleviate any teacher concerns and cultivate the trust needed to support professional growth. Equally important is the need to ensure the privacy of students and to communicate this information to parents. For more information on how to create a "safe" environment that promotes professional learning, see https://cepr.harvard.edu/files/cepr/files/2._cultivating_trust_in_video_observations.pdf .

In summary, before video recording, we strongly encourage you to set clear goals for what you hope to capture on video and its intended audience. When used well, we believe video is an essential tool for helping researchers and practitioners better understand the qualities and conditions needed to teach effectively and then, use that knowledge to improve professional practice.

References

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

Atteberry, A., Loeb, S. L., & Wyckoff, J. (2017). Teacher churning within schools: Impacts on student achievement. *Educational Evaluation and Policy Analysis, 39*(1), 3-30.

Ball, D. L., & Cohen, D. K. (1999). Developing practice, developing practitioners: Toward a practice-based theory of professional education. *Teaching as the Learning Profession: Handbook of Policy and Practice*, *1*, 3-22.

Ball, D.L., Thames, M.H., Phelps, G.C. (2008). Content Knowledge for Teaching: What Makes It Special? *Journal of Teacher Education*. 59(5) 389-407.

Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., Pianta, R. C., & Qi, Y. (2012). An argument approach to observation protocol validity. *Educational Assessment*, *17*(2-3), 62-87.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 7-74.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, *21*(1), 5-31.

Blazar, D. (2015a). Grade assignments and the teacher pipeline: A low-cost lever to improve student achievement?. *Educational Researcher*, *44*(4), 213-227.

Blazar, D. (2015b). Effective teaching in elementary mathematics: Identifying classroom practices that support student achievement. *Economics of Education Review*, 48, 16-29.

Blazar, D. (2016). What does it mean to be ranked a ''high'' or ''low'' value-added teacher? Observing differences in instructional quality across districts. *American Educational Research Journal, 53*(2), 324–359. doi:10.3102/0002831216630407

Blazar, D., & Kraft, M. A. (2017). Teacher and teaching effects on students' attitudes and behaviors. *Educational Evaluation and Policy Analysis, 39*(1), 146-170.

Blazar, D., Litke, E., & Barmore, J. (2016). What does it mean to be ranked a ''high'' or ''low'' value-added teacher? Observing differences in instructional quality across districts. *American Educational Research Journal*, 53(2), 324–359.

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. doi:10.1016/j.tate.2006.11.012

Bridges, E. M. (1990). Evaluation for tenure and dismissal. The new handbook of teacher evaluation: Assessing elementary and secondary school teachers, 147-157.

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

Cantrell, S., & Kane, T. J. (2013). Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET Project's three-year study. MET Project Research Paper.

Carpenter, T. P., Fennema, E., Peterson, P. L., Chiang, C. P., & Loef, M. (1989). Using knowledge of children's mathematics thinking in classroom teaching: An experimental study. *American Educational Research Journal*, *26*(4), 499-531.

Chambers, J., Brodziak de los Reyes, I., & O'Neil, C. (2013). How much are districts spending to implement teacher evaluation systems? RAND.

Chin, M., & Goldhaber, D. (2015). *Exploring explanations for the "weak" relationship between value added and observation-based measures of teacher performance*. Working Paper. Cambridge, MA: National Center for Teacher Effectiveness. Retrieved from: http://cepr.harvard.edu/files/cepr/files/sree2015_simulation_working_paper.pdf.

Danielson, C., & McGreal, T. L. (2000). *Teacher evaluation to enhance professional practice.* Alexandria, VA: Association for Supervision and Curriculum Development.

Darling-Hammond, L. (2013). Getting teacher evaluation right: What really matters for effectiveness and improvement? New York: Teachers College Press.

Darling-Hammond, L., Wise, A. E., & Pease, S. R. (1983). Teacher evaluation in the organizational context: A review of the literature. *Review of Educational Research*, *53*(3), 285-328.

Fennema, E., Franke, M. L., Carpenter, T. P., & Carey, D. A. (1993). Using children's mathematical knowledge in instruction. *American Educational Research Journal*, *30*(3), 555-583.

Gallup, G. H. The eleventh annual Gallup poll of the public's attitudes toward the public schools. *Phi Delta Kappan, 1979, 60,* 33-45.

Garet, M. S., Cronen, S., Eaton, M., Kurki, A., Ludwig, M., Jones, W…. & Sztejnberg, L. (2008). The impact of two professional development interventions on early reading instruction and achievement (NCEE 2008-4030). Washington, D.C.: *National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.* Retrieved from http://ies.ed.gov/ncee

Garet, M. S., Wayne, A., Stancavage, F., Taylor, J., Eaton, M., Walters, K…. & Doolittle, F. (2011). Middle school mathematics professional development impact study: Findings after the second year of implementation (NCEE 2011-4024).. Washington, DC: *National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.* Retrieved from http://ncee.ed.gov

Garet, M. S., Heppen, J.B., Walters, K., Parkinson, J., Smith, T.M., ..., Wei, T.E. (2016, November). *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.*

Glazerman, S., Isenberg, E., Dolfin, S., Bleeker, M., Johnson, A., Grider, M., & Jacobus, M. (2010). *Impacts of comprehensive teacher induction: Final results from a randomized controlled study* (No. 691d9603eb074051b57684e4affae4d4). Mathematica Policy Research.

Goe, L., Bell, C., & Little, O. (2008). Approaches to Evaluating Teacher Effectiveness: A Research Synthesis. National Comprehensive Center for Teacher Quality.

Grissom, J. A., Loeb, S., & Master, B. (2013). Effective instructional time use for school leaders: Longitudinal evidence from observations of principals. *Educational Researcher*, *42*(8), 433-444.

Gudridge, B. M. (1980). *Teacher competency: Problems and solutions.* Arlington, Va.: American Association of School Administrators

Hallgren, K., James-Burdumy, S., & Perez-Johnson, I. (2014). *State Requirements for Teacher Evaluation Policies Promoted by Race to the Top.* Mathematica Policy Research.

Hanushek, E. A. (2009). Teacher deselection. In D. Goldhaber & J. Hannaway, (Eds.), *Creating a new teaching profession* (pp. 109-118). Washington D.C.: Urban Institute Press.

Hiebert, J., & Stigler, J. W. (2000). A proposal for improving classroom teaching: Lessons from the TIMSS video study. *The Elementary School Journal*, 101(1), 3-20.

Herman J. L., & Baker, E. L. (2009). Assessment policy: Making sense of the Babel. In G. Sykes, B. Schneider, and D. N. Plank (Eds.), *Handbook of educational policy research* (pp. 176-190). New York: Routledge.

Hill, H. C. (2007). Mathematical knowledge of middle school teachers: Implications for the No Child Left Behind policy initiative. *Educational Evaluation and Policy Analysis*, 29(2), 95-114.

Hill, H. C., Blazar, & Lynch, K. (2015). Resources for teaching: Examining personal and institutional predictors of high-quality instruction. *AERA Open, 1*(4), 1-23.

Hill, H. C., Blunk, M. L., Charalambous, C. Y., Lewis, J. M., 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(4), 430-511.

Hill, H. C., Charalambous, C. Y., Author, McGinn, D., Kraft, M. A., Beisiegel, M., ... & Lynch, K. (2012). Validating arguments for observational instruments: Attending to multiple sources of variation. *Educational Assessment, 17*(2-3), 88-106.

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: teacher observation systems and a case for the generalizability study. *Educational Researcher*, 41(2), 56-64.

Hill, H. C., & Grossman, P. (2013). Learning from teacher observations: Challenges and opportunities posed by new teacher evaluation systems. *Harvard Educational Review, 83*(2), 371-384.

Hill, H. C., Schilling, S. G., & Ball, D. L. (2004). Developing of teachers' measures mathematics knowledge for teaching. *Elementary School Journal*, 105(1), 11-30. doi: 10.1086/428763

Jackson, C. K., Rockoff, J. E., & Staiger, D. O. (2014). Teacher effects and teacher-related policies. *Annual Review of Economics*, 6(1), 801-825.

Jacobs, S., Doherty, K. M., Lakis, K., Lasser, P., & Staresina, L. N., (2014). *2015 state teacher policy yearbook.* Washington, DC: National Center on Teacher Quality.

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.* doi: 10.1080/19345747.2016.1273411

Kane, T., & Staiger, D. O. (2012). Gathering feedback on teaching: Combining high-quality observations with student surveys and achievement gains. Seattle, WA: Bill & Melinda Gates Foundation.

Kane, T., Gehlbach, H., Greenberg, M., Quinn, D., & Thal, D. (2015). The Best Foot Forward Project: Substituting teacher-collected video for in-person classroom observations. First year implementation report. Cambridge, MA: Center for Education Policy Research, Harvard University. Retrieved from: https://cepr.harvard.edu/files/cepr/files/l4a_best_foot_forward_research_brief1.pdf

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). Have we identified effective teachers? Validating measures of effective teaching using random assignment. In *Research Paper. MET Project. Bill & Melinda Gates Foundation*.

Kazemi, E., & Franke, M. L. (2004). Teacher learning in mathematics: Using student work to promote collective inquiry. *Journal of Mathematics Teacher Education*, 7(3), 203-235.

Kraft, M.A. & Blazar, D. (2014). Improving Teachers' Practice across Grades and Subjects: Experimental Evidence on Individualized Coaching. *Providence, RI: Brown University.*

Kraft, M.A., Blazar, D., Hogan, D. (2016). The effect of teaching coaching on instruction and achievement: A meta-analysis of the causal evidence. Brown University Working Paper. Retrieved from: https://scholar.harvard.edu/files/mkraft/files/kraft_blazar_hogan_2016_teacher_coaching_meta-analysis_wp_w_appendix.pdf

Kraft, M. A., & Gilmour, A. F. (2016). Can principals promote teacher development as evaluators? A case study of principals' views and experiences. *Educational Administration Quarterly*, 52(5), 711-753.

Kraft, M. A., & Gilmour, A. F. (2017). Revisiting the widget effect: Teacher evaluation reforms and the distribution of teacher effectiveness. *Educational Researcher*, 46(5), 234-249.

Lamon, S. J. (2012). Teaching fractions and ratios for understanding: Essential content knowledge and instructional strategies for teachers. Routledge.

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

May, H., & Supovitz, J. A. (2011). The scope of principal efforts to improve instruction. *Educational Administration Quarterly*, 47(2), 332-352.

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics.* Reston, VA.

National Council of Teachers of Mathematics. (1991). *Professional standards for teaching mathematics.* Reston, VA.

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

National Council on Teacher Quality (2013). *2013 State Teacher Policy Yearbook*. Washington, DC: Author.

National Governors Association Center for Best Practices, Council of Chief State School Officers. (2010). *Common core state standards for mathematics*. Washington, D.C.

Odden, A. (2004). Lessons learned about standards-based teacher evaluation systems. *Peabody Journal of Education, 79*(4), 126–137.

Ost, B. (2014). How do teachers improve? The relative importance of specific and general human capital. *American Economic Journal. Applied Economics*, *6*(2), 127.

Papay, J. P. (2012). Refocusing the debate: Assessing the purposes and tools of teacher evaluation. *Harvard Educational Review, 82*(1), 123-141.

Partee, G. L. (2012). Using Multiple Evaluation Measures to Improve Teacher Effectiveness: State Strategies from Round 2 of No Child Left Behind Act Waivers. Center for American Progress.

Range, B., Scherz, S., Holt, C., & Young, S. (2011). Supervision and evaluation: the Wyoming perspective. *Educational Assessment, Evaluation and Accountability, 23*(3), 243–265.

Santagata, R. (2009). Designing Video-Based Professional Development for Mathematics Teachers in Low-Performing Schools. *Journal of Teacher Education*, 60(1), 38-51. doi:10.1177/0022487108328485

Sherin, M. G., & van Es, E. A. (2005). Using video to support teachers' ability to notice classroom interactions. *Journal of technology and teacher education*, 13(3), 475.

Stigler, J. W., Gallimore, R., & Hiebert, J. (2000). Using video surveys to compare classrooms and teaching across cultures: Examples and lessons from the TIMSS video studies. *Educational Psychologist*, 35(2), 87-100.

Stigler, J. W., Gonzales, P., Kawanaka, T., Knoll, S., & Serrano, A. (1999). The TIMSS videotape classroom study: Methods and findings from an exploratory research project on eighth grade mathematics instruction in Germany, Japan, and the United States. Retrieved from Washington, D.C.: https://nces.ed.gov/timss/

Stigler, J. W., & Hiebert, J. (1997). Understanding and improving classroom mathematics instruction: An overview of the TIMSS video study. *Phi Delta Kappan*, *79*(1), 14.

Taylor, E. S., & Tyler, J. H. (2012). The effect of evaluation on teacher performance. *The American Economic Review*, *102*(7), 3628-3651.U.S. Department of Education. Race to the Top Fund. https://www2.ed.gov/programs/racetothetop/awards.html

U.S. Dept. of Education, Race to the Top Fund, Retrieved from: https://www2.ed.gov/programs/racetothetop/index.html

Weisberg, D., Sexton, S., Mulhern, J., Keeling, D., Schunck, J., Palcisco, A., & Morgan, K. (2009). T*he widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness*. The New Teacher Project.

Wright, S., Horn, S., & Sanders, P. (1997). Classroom context effects on student achievement: implications for teacher evaluation. Journal of Personnel Evaluation in Education, 11, 57–67.

Yoon, K. S., Duncan, T., Lee, S. W. Y., Scarloss, B., & Shapley, K. L. (2007). Reviewing the Evidence on How Teacher Professional Development Affects Student Achievement. Issues & Answers. REL 2007-No. 033. Regional Educational Laboratory Southwest (NJ1).

Author Information

David Blazar
University of Maryland
Dept. of Teaching & Learning, Policy & Leadership
2311E Benjamin Bldg., 3942 Campus Dr.
College Park, MD 20742
USA

Barbara Gilbert
Harvard University
Center for Education Policy Research
50 Church St., 4th Floor
Cambridge, MA 02138
USA
Contact e-mail: *barbara_gilbert@gse.harvard.edu*

Corinne Herlihy
Harvard University
Center for Education Policy Research
50 Church St., 4th Floor
Cambridge, MA 02138
USA

Claire Gogolen
Harvard University
Center for Education Policy Research
50 Church St., 4th Floor
Cambridge, MA 02138
USA