CAN VIDEO TECHNOLOGY IMPROVE TEACHER EVALUATIONS? AN EXPERIMENTAL STUDY

Thomas J. Kane
Harvard Graduate School of Education
Cambridge, MA 02138
tom_kane@gse.harvard.edu

David Blazar
(corresponding author)
College of Education
University of Maryland
College Park, MD 20742
dblazar@umd.edu

Hunter Gehlbach
School of Education
Johns Hopkins University
2800 N. Charles St.
Baltimore, MD 21218
gehlbach@jhu.edu

Miriam Greenberg
Center for Education Policy Research
Harvard Graduate School of Education
Cambridge, MA 02138
miriam_greenberg@gse.harvard.edu

David M. Quinn
Rossier School of Education
University of Southern California
Los Angeles, CA 90089
quinnd@usc.edu

Daniel Thal
Mathematica
Cambridge, MA 02139
dthal@mathematica-mpr.com

Abstract

Teacher evaluation reform has been among the most controversial education reforms in recent years. It also is one of the costliest in terms of the time teachers and principals must spend on classroom observations. We conducted a randomized field trial at four sites to evaluate whether substituting teacher-collected videos for in-person observations could improve the value of teacher observations for teachers, administrators, or students. Relative to teachers in the control group who participated in standard in-person observations, teachers in the video-based treatment group reported that post-observation meetings were more "supportive" and they were more able to identify a specific practice they changed afterward. Treatment principals were able to shift their observation work to noninstructional times. The program also substantially increased teacher retention. Nevertheless, the intervention did not improve students' academic achievement or self-reported classroom experiences, either in the year of the intervention or for the next cohort of students. Following from the literature on observation and feedback cycles in low-stakes settings, we hypothesize that to improve student outcomes schools may need to pair video feedback with more specific supports for desired changes in practice.

https://doi.org/10.1162/edfp_a_00289

1. INTRODUCTION

Citing evidence of large differences in student achievement gains between individual teachers' classrooms (Hanushek and Rivkin 2010), the Obama administration incentivized states to redesign their teacher evaluation systems through the Race to the Top program in 2009 and through their approval of state plans under the No Child Left Behind Act. Yet, the evidence of the success of those efforts has been mixed. In Washington DC, Chicago, and Newark, high-stakes teacher evaluations seemed to lower the retention rates of low-performing teachers and increase the retention of more effective teachers (Dee and Wyckoff 2015; Fulbeck et al. 2016; Sartain and Steinberg 2016), both desirable outcomes. In Chicago and Cincinnati, the feedback seemed to improve the practice of existing teachers (Taylor and Tyler 2012; Steinberg and Sartain 2015). However, in other states, many have judged such payoffs insufficient to justify the cost in terms of political controversy, teacher and principal time, and the ability to recruit new and high-quality teachers (Jiang, Sporte, and Luppescu 2015; Kraft et al. 2019; Stecher et al. 2019). Although the laws remain on the books, some state agencies have deemphasized teacher evaluation following the passage of the Every Student Succeeds Act in 2015 (Sawchuk 2016).

Teacher evaluations typically include two main components: test-based measures of student achievement growth and classroom observations by a school administrator. Although the test-based measures tend to generate the greatest political controversy (Ballou and Springer 2015; Jiang, Sporte, and Luppescu 2015), the costliest component, in terms of principal and teacher time, is the classroom observation. According to Dynarski (2016) and our own surveys, supervisors spend between ten and thirty hours for each teacher performing observations, writing their comments, and discussing the results with teachers. When multiplied across 3.1 million public school teachers, at the average principal's salary of roughly $45 per hour (USDOE 2012; Dynarski 2016), the cost of in-person observations would be between $1.4 and $4.2 billion per year. These large estimates also do not account for additional social costs, including stress on both principals and teachers (Grissom, Loeb, and Master 2013).

Given the time devoted to classroom observations, our goal was to test whether the substitution of teacher-collected video for in-person observation could improve the value of the evaluation process for teachers, administrators, and students. We hypothesized that digital video would offer several advantages over in-person observations: Video provides a more detailed, third-party record for teachers and principals to discuss; watching videos of their own instruction may be more revelatory for teachers than an observer's written notes; giving teachers control of the camera elevates the role of teachers in their own evaluations; video allows principals to time-shift their observational duties to quieter times of the day or week; and video makes it feasible to incorporate the perspective of external observers and content experts.

If proven effective, the purchase of video-based technology would be a relatively inexpensive way to increase the value of teacher observations. We estimate the cost of the program we evaluated to be roughly $2,500 per teacher, which includes: (1) the cost of tablets and stands (ranging from $500 to $1,000 per unit); (2) computer hardware, software, storage, and IT support (roughly $1,500 per teacher); and (3) feedback from

outside content experts (roughly $250 per teacher).[1] Costs likely would be substantially lower in future years, given economies of scale for this sort of technology: Videos and tablets can be shared across teachers, and reused across school years; and the marginal cost for ongoing use or for adding an additional teacher is much lower than the baseline costs for other hardware and software.

However, watching video of oneself can be unpleasant and anxiety-producing (Raymond, Dorwick, and Kleinke 1993). Thus, to expand beyond voluntary adopters, teachers will need a reason and incentive to do so. The schools in our study offered teachers a trade: In return for teachers' willingness to use video for classroom observations, teachers would control the camera and choose which lesson videos to submit for their formal observations. A secure software platform allowed observers, including both formal evaluators and content experts, to watch the videos and provide time-stamped comments aligned to specific moments in the videos. These videos and comments were used in one-on-one discussions between teachers and principals and external content experts. To test the efficacy of such a system, we conducted a randomized field trial involving 433 teachers and 134 school administrators at four different sites in Delaware, Georgia, Colorado, and California.

We found that combining the cameras and the ability to substitute video for in-person observations did shift the way teacher evaluations were conducted. In the first year of the study, the average teacher collected thirteen videos of her practice, rather than the three in-person observations generally required for formal evaluation procedures. Following their observations, teachers in the treatment group were more likely than those in the control group to report that their post-observation conversations with supervisors were "supportive" and their observations were "fair" (language comes directly from the teacher survey). Teachers in the treatment group also were more likely to identify specific practices they changed after being observed by and meeting with their principal. While video-based observations did not save time for principals in the aggregate, principals spent less time on paperwork and more time observing and interacting directly with teachers.

We also found that treatment teachers were substantially more likely to remain in their school in the year following the intervention. Differences in retention rates around 10 percentage points are larger than many other educational interventions. Our randomized design does not allow us to directly test the mechanisms driving these retention effects. However, given additional evidence on the role of teacher–principal relationships in teacher departures (Boyd et al. 2011; Kraft, Marinell, and Yee 2016), we see the large retention effects as consistent with the impacts we observed on teacher perceptions of supervisors' supportiveness and fairness.

Ultimately, though, the intervention did not produce measurable differences in student perceptions of classroom instruction or improved student performance on state tests in math and reading. We found null effects both at the end of the intervention year

1. To estimate costs per teacher, we calculated the total costs paid for this program and divided by the number of treatment participants. Total computer costs including hardware, software, storage, and IT support were roughly $1.3 million for roughly 215 teachers; total costs for outside content experts were roughly $215,000 for the same number of teachers.

and, for the first cohort of teachers, for the students they taught in the year following the intervention. Given the more favorable literature on the impacts of teacher coaching (Kraft, Blazar, and Hogan 2018)—which also relies on observation of teachers' practice but in a low-stakes environment—we hypothesize that use of video for teacher evaluation may need to be paired with direct feedback on, and practice related to, specific instructional behaviors in order to generate changes in student outcomes.

2. LITERATURE REVIEW

Although the theory of action linking classroom observations to improved student outcomes is often unstated, it supposes that (1) observation rubrics can identify instructional behaviors that are related to student outcomes; and (2) such rubrics provide a common vocabulary teachers and supervisors can share for discussing key aspects of instruction; so that (3) an observer's written and oral feedback during the post-observation conference will lead the teacher to recognize previously unrecognized aspects of her or his behavior that fall short of the standards, (4) the conversation between the principal and teacher will lead the teacher to identify the instructional changes she should make to improve on the standards, and (5) despite the evidence on the difficulty of adult behavior change, the teacher will be able to incorporate the new behaviors in her instruction; and ultimately (6) student achievement will rise as a result of the improved teacher behaviors.

There is evidence to support several of these propositions. For instance, observation scores on several of the major observational rubrics have been shown to be correlated with student achievement gains (Kane et al. 2013; Araujo et al. 2014; Blazar 2015a). Moreover, a number of studies have confirmed that when observers are trained on one of the major observational rubrics, they can apply them reliably—although achieving a reliability coefficient greater than 0.7 requires averaging over several adult observers and several lessons (Bell et al. 2012; Hill, Charalambous, and Kraft 2012; Kane and Staiger 2012). Whereas most prior research has relied on trained raters to score lessons by teachers they do not know, principals have been shown able to score their own teachers' videos as reliably as principals from other schools, albeit with an upward shift in mean scores (Ho and Kane 2013).

A growing body of experimental evidence on teacher coaching indicates that using observation protocols to provide teachers with feedback on their instruction in a nonevaluative setting can help them improve their classroom performance, as well as student achievement. A recent meta-analysis of the causal evidence on the effectiveness of teacher coaching found average treatment effects of 0.49 standard deviations (SD) on observed teacher practice and 0.18 SD on student achievement (Kraft, Blazar, and Hogan 2018). Though coaching practices can result in improved performance through several possible pathways (e.g., opportunities to receive direct feedback, practice teaching skills, observe models of successful teaching), one likely mechanism is the opportunity to notice and reflect on one's own practice. In effect, the coach may serve as a mirror with which to see one's own practice—a role that digital video also could play in a higher-stakes evaluation setting. Descriptive studies have found an association between teacher observations of videos and changes in practice (Brunvand and Fishman 2006; Rosaen et al. 2008; Santagata and Angelici 2010; Kleinknecht and Schneider 2013).

Some researchers and practitioners have raised concerns about conflating coaching interventions, which are purposefully low-stakes and non-evaluative, with official teacher evaluation, which could carry consequences for the teachers' employment, earnings, or daily work relationships (Kraft and Gilmour 2016). On one hand, the absence of formal consequences to coaching interventions may lower teachers' anxiety and make them more receptive to feedback. On the other hand, the incentive to actually change practice may be weaker when there are no stakes attached.

Our review of a handful of teacher evaluation studies suggests that, under certain conditions, both types of observation-based interventions can influence teacher behavior and student outcomes. For example, evidence from the IMPACT program in Washington, DC, indicates that combining teacher evaluations with financial incentives and dismissal threats (Dee and Wyckoff 2015) led to higher exit rates for low-performing teachers and some improvements in teacher practice for middle- and high-performing teachers. In addition, Taylor and Tyler (2012) studied the impact of implementing a formal, rubric-based classroom observation for experienced teachers in Cincinnati (Ohio) Public Schools between 2005 and 2010. Experienced teachers were evaluated every five years, based on their hire date. During their evaluation year, teachers were observed four times (three times by a trained observer from outside their school and once by their supervisor or principal). After each classroom observation, the observers provided written comments to the teacher and they met at least once in person. Controlling for student baseline scores and characteristics, the authors found that student achievement rose 0.07 SD during the evaluation year and remained 0.11 SD higher in the year after evaluation. Similarly, teachers in a pilot program in Chicago were evaluated multiple times per year using the Danielson Framework for Teaching (Danielson 2011) observation instrument. Teachers who participated in the pilot evaluation system had higher student-achievement in reading of 0.10 SD (Steinberg and Sartain 2015).

3. HYPOTHESES

We hypothesized that the introduction of video would improve the evaluation process in five ways. First, the traditional in-person observation as practiced in U.S. schools—in which a supervisor observes, takes notes, and presents written feedback to teachers—may unnecessarily add to the areas of conflict between a teacher and supervisor (Hill and Grossman 2013; Jiang, Sporte, and Luppescu 2015; Kraft and Gilmour 2016). Although supervisor–employee relationships inherently involve some tension, the designation of the supervisor as note-taker unnecessarily invites disputes over the facts. Supervisors both control the official record of the teacher's and students' behaviors during the observation (in the form of their notes), as well as the interpretation of those facts. There is almost an infinite number of pieces of data generated amidst the interactions between a teacher and students over the course of a lesson. A teacher is noticing and remembering only a subset; and the observer is noticing another—potentially nonoverlapping—subset. Although there may be disputes over the *interpretation* of what occurred in the lesson, the recording of the video essentially eliminates potential disputes over the *facts* of what happened during a lesson.

A second potential benefit of video is providing teachers with opportunities to supplement their own recollection of a lesson by watching it again from another vantage point. In watching the video, they may notice behaviors that they had not noticed in

real time, given the limits on working memory and peripheral vision. They may also notice behaviors the observer also failed to record in her or his notes. Moreover, in those instances where an observer's notes do accurately record behaviors the teacher did not notice in real time, the video may warrant a higher level of veracity than the supervisors' notes. Thus, with a more complete and accurate set of data around the facts of a lesson, a teacher would have more opportunities to recognize behaviors she wants to change. Relatedly, a third way that video may change the observation process is, once a teacher has identified a behavior she wants to change, she may be more able to practice alternative behaviors and, thus, verify her success by recording and viewing subsequent lessons.

A fourth potential benefit is that the ability to share video electronically lowers the cost of engaging observers—especially those outside the school—with expertise in a teacher's content area. (See, for example, a growing literature on use of video for teacher coaching in order to leverage outside expertise; Allen et al. 2011.) Evaluation requires identifying content-and grade-level experts (Hill and Grossman 2013), which can be a challenge in practice from teacher and principal perspectives (Kraft and Gilmour 2016).

A fifth way in which we hypothesized our treatment to improve the evaluation process was the ability of teachers to select lessons ex post. The schools in our sample typically required observers to give teachers 24 hours' notice before an in-person observation, thus allowing a teacher to better prepare the lesson to be observed. A teacher can prepare herself ex ante, but she is still subject to the risk that the lesson does not go as planned. In our intervention, teachers could reduce their exposure to the risk of in-class surprises by choosing to submit only those lessons they perceived to have gone well. Of course, this process could generate both benefits and costs. The reduction in teacher anxiety resulting from ex post lesson selection could improve workplace relationships. However, it could also make classroom observations less informative for supervisors, if it were to obscure poor teaching practice. Yet, when other teachers have the same opportunity to choose lessons, teacher-selection of lessons need not prevent supervisors from identifying their weakest teachers. The best lessons from the best teachers might remain higher than the best lessons from the weakest teachers. (In the following, we present evidence that teacher selection of lessons largely preserved the rankings on teacher observation scores.)

Video-based observation could also be crowding out unrelated class preparation and instructional activities. The time spent planning the lessons to be recorded or viewing the lessons afterwards could diminish time spent by teachers in preparing for unrecorded lessons.

4. METHODS

To assess the benefits of video-based teacher evaluation, in the spring of 2013, the study team recruited principals at four sites: small districts across the state of Delaware, a midsized district in Georgia, a collection of smaller districts in Colorado, and a large district in California. Project staff first recruited schools to participate in a test of video-based evaluations, and then worked with school leaders to recruit teachers. For a school to be eligible, a minimum of three teachers in a school must have agreed to participate in the study. In recruitment sessions, all teachers in relevant grades were invited to participate. After the initial recruitment sessions, project staff sent materials to principals

who then forwarded these to teachers. All materials framed participation as a "voluntary opportunity" to ensure teachers did not perceive that principals were coercing them to participate.

In October 2013, eligible schools (and the teachers in each who agreed to participate) were randomly assigned to the treatment or control group. This process was repeated again in 2014, when a second cohort of schools was recruited from the large California school district. The total randomized sample consisted of 134 school administrators and 433 teachers. Random assignment occurred at the school level within each of the four study sites, with 52 schools randomly assigned to treatment and 55 schools randomly assigned to control. There were 85 schools in cohort 1 ($N = 345$ teachers and 107 administrators) and 22 in cohort 2 ($N = 88$ teachers and 27 administrators).[2] Of the participating teachers, 54 percent were in upper-elementary grades (i.e., grade 4 or 5), and 46 percent were in middle school (i.e., grades 6 through 8).

While teachers in the control group continued with their traditional in-person observation process, teachers in the treatment group participated in a multifaceted intervention designed to test the value of video-based observation and evaluation. First, teachers were given a video camera with which to record their own lessons. A private contractor, BloomBoard, provided video storage and a software platform for teachers to collect a library of videotaped lessons and observation artifacts (such as lesson plans and handouts). Working with a hardware supplier, thereNow, the study team distributed camera kits to all treatment teachers. 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. When the device was plugged into an Ethernet port, the file was piped securely to a teacher's individual online account. Each teacher had a unique log-in, and only she could view and share videos in her account. Teachers who joined the project in the second year used the Swivl video recording device and two microphones, both of which attached to an iPad mini. In both years, teachers chose which of these videos they uploaded to Bloomboard from their device.[3] Teachers were asked to record two lessons per month and upload all lessons to the secure server. Teachers chose three videos to submit for their formal evaluation, and two videos for viewing by nonevaluative feedback from content experts outside the school.

After a teacher shared a video with an observer, the observer logged in, tagged specific moments of the video, and commented on specific moments in the lesson. The software was customized so that the tags would correspond to each district's observation rubric. Rubrics varied by district or state, but included many similar components

2. One school participated in both cohorts, as a control school in cohort 1 and a treatment school in cohort 2.
3. In the spring prior to the start of the school year in which the intervention took place, participants in the treatment group were trained to use the platform and video cameras for their observations. The training consisted of three to four hours of hands-on workshop-style activities. The team visited each site for camera distribution and training, and received ongoing training and technology support. The training included guidance for administrators on methods for giving feedback using video evidence. The training focused on minimizing teacher-perceived vulnerability, focusing on high-leverage moments in the video and using questioning strategies to shift the analysis of practice from administrator to teacher.

(e.g., planning and preparation for instruction, instructional delivery, classroom environment, professional responsibilities) that align closely to widely utilized instruments, such as the Danielson Framework (Danielson 2011). 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 they met in person to discuss the video feedback and determine a final score.[4]

Most treatment teachers also received nonevaluative feedback from coaches provided by a nonprofit contractor, TNTP (formerly The New Teacher project). TNTP assigned teachers a coach based on content area (i.e., elementary education, math, or English language arts [ELA]). We asked teachers to share two videos with their assigned coach: the first in the fall (October and November) and the second in the winter (January and February) of each school year. Coaches viewed the videos on the BloomBoard platform and added written comments within one week of upload. We encouraged teachers to debrief with their coach via phone following each observation, though we do not have a record of the number or content of the phone conversations. Consistent with program guidelines, 76 percent of teachers completed both virtual coaching sessions, and 96 percent completed one of the two. We are not able to disentangle the coaching component of the intervention from the principal-based evaluation. However, the intensity of coaching in this intervention was lower (two sessions) than many other interventions focused solely on coaching, which often include several week-long observation and feedback cycles (Kraft, Blazar, and Hogan 2018).

Data Collection

Throughout the intervention, the research team collected a variety of sources of data on participants in both the treatment and control groups. Teachers and principals completed a baseline survey asking about their teaching experience and prior experiences with classroom observations. In the first year of the study, we asked teachers and principals to complete a post-observation survey in which they reflected on their experiences with this process. We also surveyed principals weekly from November through May of the intervention year regarding time spent on teacher observation activities. (The weekly survey data were not collected in cohort 2.) In both cohorts, we also surveyed teachers and principals at the end of the school year about their overall experience with the evaluation and observation process. In the analyses presented below, we conduct analyses for individual survey items and, thus, do not present reliability indices for these measures.

The project team also administered a survey to students at the end of each school year. Survey items ($N = 24$) assessed the extent to which students experienced the classroom environment as engaging, demanding, and supportive of their intellectual growth.[5] Exploratory factor analyses indicated two factors with an eigenvalue above 1.0 (Kline 1994); scree-plot analysis also supports this two-factor solution (Hayton, Allen,

4. Many teachers also received developmental feedback (which did not contribute to their formal evaluation) on two of their recorded lessons from a virtual coach provided by TNTP. This component of the intervention was voluntary.

5. The survey instrument was developed by Hunter Gehlbach, informed by the constructs from Tripod most highly correlated with student achievement (Kane and Cantrell 2010).

and Scarpello 2004). The first factor consists of all items (alpha = 0.89) and that we consider to measure students' overall classroom experiences. The second factor consists of seven items (alpha = 0.77) focused on students' classroom behavior and teachers' ability to manage (mis)behavior in class. These data were available for both cohorts at the end of the treatment year, and for a subset of cohort 1 teachers at the end of the follow-up year.

Finally, we assembled administrative data on student characteristics and achievement from the participating districts. These data included demographic information on students (i.e., gender, race/ethnicity, free or reduced-price lunch [FRPL] eligibility, those in need of an individualized education plan [IEP], and limited English proficiency [LEP] status), as well as current- and prior-year test scores in math and ELA on state assessments. We standardized test scores within districts by grade, subject, and year using the entire population of students. After the intervention had recruited schools, the state of California announced a statewide hiatus in testing for the spring of 2014, as they piloted a new Common Core–aligned assessment, so the project did not have student achievement data for the California schools for the first year. In the first year of the study, administrative records also included formal evaluation scores for teachers.

Further, we used administrative records from the districts in order to examine turnover of teachers. As administrative data only were collected through 2015, our retention analyses focus on the first cohort, whom we could observe in the follow-up year. We measured retention in three ways: whether teachers maintained their teaching assignment in the year following the intervention, in the same school and grade; whether teachers stayed in their same school but taught a new grade level;[6] and whether teachers remained in their district but moved to a different school. The remaining teachers were not observed in administrative records in the follow-up year. We infer that these teachers left the district or teaching altogether.[7]

External Validity

Our goal was to inform the design and implementation of teacher evaluation systems across the United States. Although our sample consists of volunteers, participants look similar to others in their respective schools and districts in terms of student and teacher observables. As reported in table 1, participating students and teachers were similar to nonparticipants. In column 2, we compare participating teachers and their students to nonparticipating teachers and students within the same school (i.e., including school fixed effects), given that the school was the level of randomization. The participating classrooms had a slightly higher percentage of FRPL-eligible students and a slightly lower percentage of students with IEPs. In column 3, we make comparisons between schools participating in the experiment and other schools in the districts. There were no

6. We differentiate between retained in school and grade versus retained in school, given research indicating that switching grades from one year to the next is negatively associated with gains in students' academic performance (Ost 2014; Blazar 2015b; Atteberry, Loeb, and Wyckoff 2017).

7. It is possible that some of the teachers who were not observed in administrative records in the year following the intervention moved teaching assignments in a way that made them unobservable in these data. For example, in three districts we only had these records for elementary and middle schools. Therefore, it is possible that teachers may have switched to teaching high school. We believe that both types of moves are of substantive interest.

Table 1. External Validity: Study Participants versus Nonparticipants

	Participating Schools		
	Study Participants	Participating Classes — Nonparticipating Classes Difference (SE)	Participating Schools — Nonparticipating Schools Difference (SE)
Student characteristics			
Proportion male	0.511	−0.003 (0.005)	−0.001 (0.003)
Proportion FRPL-eligible	0.577	0.031** (0.014)	−0.011 (0.043)
Proportion with IEP	0.102	−0.036*** (0.013)	0.008 (0.005)
Proportion designated LEP	0.264	−0.007 (0.004)	0.007 (0.034)
Average prior score: Math	0.077	0.044 (0.041)	0.009 (0.060)
Average prior score: ELA	0.057	0.052 (0.035)	0.004 (0.065)
Proportion African American	0.186	0.000 (0.004)	−0.008 (0.016)
Proportion Asian	0.063	−0.003 (0.003)	0.006 (0.014)
Proportion Hispanic	0.385	−0.006 (0.005)	−0.008 (0.040)
Proportion Native American	0.007	0.000 (0.000)	0.000 (0.001)
Proportion Pacific Islander	0.001	0.000 (0.000)	−0.000 (0.001)
Proportion white	0.339	0.009 (0.005)	0.009 (0.029)
Proportion Multiple/Other Race	0.019	−0.000 (0.001)	0.001 (0.001)
N (Students)	22,950		
Teacher characteristics			
Proportion male	0.291	−0.026 (0.041)	−0.025 (0.022)
Average years of teaching experience	10.330	−0.107 (0.502)	0.148 (0.176)
Proportion African American	0.087	−0.022 (0.018)	−0.002 (0.020)
Proportion Hispanic	0.180	0.000 (0.025)	0.013 (0.038)
Proportion white	0.669	0.021 (0.030)	−0.006 (0.036)
N (Teachers)	426		

Notes: The student sample excludes special education classes (defined as classes where 75 percent or more of students have an individualized education plan, or IEP) taught by non-project teachers. The sample also excludes students in treatment teachers' classes who did not have administrative data ($N = 87$; see table 3 for description of missing data). Prior scores are reported in standard deviation units, after standardizing scores by state, grade, and subject. The difference between treatment teachers and non-sample teachers in participating schools (column 2) was estimated controlling for school fixed effects. The difference between students and teachers in participating and nonparticipating schools (column 3) was estimated after controlling for district fixed effects. In all cases, standard errors (SE) are reported in parentheses, and allow for clustering within a school. Teacher gender and race were not provided for nonsample teachers or schools in the Georgia district and one of the Colorado districts, so those sites are excluded for those rows. All sites provided experience for all teachers. FRPL = free or reduced-price lunch; LEP = limited English proficiency; ELA = English language arts.

**Significant at the 95% level; ***significant at the 99% level.

differences for any of the characteristics we observe. Finally, in Appendix table A.1, we compare participating districts in Colorado and Delaware to nonparticipating districts, finding only one difference per state. We did not conduct these analyses for the sites in California or Georgia, where we had only one district per state in our sample.

Internal Validity

Table 2 summarizes the differences in baseline characteristics between the students with teachers randomly assigned to treatment or control group. With one exception, none of the differences in observed traits of administrators, teachers, or students was statistically distinguishable from zero at baseline. One exception is the percentage of Asian students. However, when differences for all characteristics are tested jointly using a Fisher-Pearson-Wald test (Young 2018), we find no difference between the two groups ($p = 0.305$).

Another threat to internal validity is differential attrition and missing data among participating administrators, teachers, and students, which could result in unbalanced groups. At the start of the experiment, 433 teachers agreed to participate. Between that time and the end of the experiment, several teachers dropped from the study for one of three reasons: they no longer wanted to participate in the intervention ($N = 10$, including 6 in the treatment group and 4 in the control group); they could not participate because they left their school, the district, or the teaching profession ($N = 10$, with equal split between treatment and control); or they participated in the intervention but did not complete/did not have their students complete end-of-year surveys ($N = 18$ for teacher surveys, with 4 from the treatment group and 14 from the control group; $N = 23$ for student surveys, with 9 from the treatment group and 14 from the control group). Of the 134 administrators who originally agreed to participate, four left their school (equal split between treatment and control), two left the study (both from the treatment group), and eleven did not complete surveys (one from treatment group and ten from control group). For test-score outcomes, we were able to capture data on most teachers even if they stopped participating in study activities. However, we are missing student test score data on all seventy-eight teachers from California in the first year of the study, given the hiatus in state testing that year (which was announced after the start of our study). As a result, we excluded the California teachers from the test-based outcomes at the end of the first year of implementation; we were able to look at their students' outcomes in the follow-up year. Of the remaining teachers, eighteen were not linked to students in the administrative data (with seven from the treatment group and eleven from the control group).

In table 3, we examine differences in the percent of participants in the treatment and control groups with each type of data. For most administrator- and teacher-level survey outcomes, we find no difference in response rates between treatment and control groups. One exception is that administrators in the treatment schools were more likely to complete the end-of-year survey. For student-level outcomes, we examine whether there were differences between treatment and control in the share of teachers who had any students who contributed to the analyses (surveys or test scores), as well as the percent of students from these teachers who had outcome data. Although we find no differences at the end of the intervention year, we do find differences in the follow-up year in the share of teachers with any students who contributed to the survey or test

Table 2. Internal Validity: Difference between Treatment and Control Groups at Baseline

	Control Mean	Treatment − Control Difference (SE)
Administrator characteristics		
Proportion male	0.397	0.113 (0.083)
Years as administrator	10.302	−1.076 (1.248)
Proportion African American	0.283	−0.056 (0.067)
Proportion Hispanic	0.200	−0.007 (0.056)
Proportion white	0.483	0.079 (0.072)
N (Administrators)		129
Fisher Pearson Wald test		$p = 0.826$
Teacher characteristics		
Proportion male	0.234	−0.023 (0.039)
Years as teacher	11.709	0.450 (0.697)
Proportion African American	0.175	0.066 (0.044)
Proportion Hispanic	0.144	−0.013 (0.034)
Proportion white	0.593	−0.023 (0.046)
N (Teachers)		426
Fisher Pearson Wald test		$p = 0.697$
Student characteristics		
Proportion male	0.509	−0.007 (0.010)
Proportion FRPL-eligible	0.590	0.035 (0.034)
Proportion with IEP	0.110	0.006 (0.016)
Proportion designated LEP	0.280	0.027 (0.021)
Average prior score: Math	0.086	−0.016 (0.073)
Average prior score: ELA	0.079	0.013 (0.063)
Proportion African American	0.164	0.005 (0.027)
Proportion Asian	0.049	−0.030** (0.013)
Proportion Hispanic	0.401	0.016 (0.023)
Proportion Native American	0.007	−0.001 (0.003)
Proportion Pacific Islander	0.001	−0.000 (0.001)
Proportion white	0.358	0.008 (0.028)
Proportion Multiple/Other race	0.020	0.003 (0.003)
N (Students)	12,759	22,950
Fisher Pearson Wald test		$p = 0.305$

Table 2. Continued.

Notes: The adjusted difference between control and treatment is the result of a regression of the dependent variable against fixed effects for randomization strata and a treatment indicator. Standard errors (SE) are reported in parentheses, and in the teacher and administrator models they allow for clustering within a school. School characteristics are from the 2012–13 school year, as they were the most recent data available at the time of randomization. FRPL = free or reduced-price lunch; IEP = individualized education plan; LEP = limited English proficiency; ELA = English language arts.

**Significant at the 95% level.

Table 3. Attrition and Missing Data

	Pooled Year 1		Cohort 1 Follow-Up Year	
	Control Mean	Treatment – Control Difference (SE)	Control Mean	Treatment – Control Difference (SE)
Administrator outcomes				
End-of-year survey	0.815	0.110** (0.051)		
Post-conference survey	0.849	0.073 (0.053)		
Time use survey	0.962	0.033 (0.022)		
Teacher outcomes				
End-of-year survey	0.888	0.040 (0.031)		
Post-conference survey	0.877	0.038 (0.038)		
Official observation score	0.788	0.058 (0.054)		
Student outcomes				
End-of-year survey: Teacher has any students with data	0.893	0.021 (0.034)	0.875	−0.181** (0.068)
End-of-year survey: Share of student surveys returned	0.755	0.039 (0.043)	0.758	−0.069 (0.069)
Test Scores: Teacher has any students with data	0.936	0.030 (0.032)	0.652	0.125** (0.050)
Test Scores: Share of students who are present that are in analysis sample	0.970	−0.010 (0.006)	0.859	0.033 (0.036)

Notes: The adjusted difference between control and treatment is the result of a regression of the dependent variable against fixed effects for randomization strata and a treatment indicator. Standard errors (SE) are reported in parentheses, and allow for clustering within-school. For student-level outcomes, samples reflect students of teachers for whom we have available data. Students whose teachers did not provide class rosters for the end-of-year survey or were not in test-score files, were not included.

**Significant at the 95% level.

score analyses. Treatment teachers were less likely to have student surveys in the follow-up year. Treatment teachers were more likely to have test score data on their students. As a result, we interpret these follow-up analyses with caution.

Data Analysis

We analyze the effect on our teacher-, principal-, and student-level outcome measures using the following specification, in which Y represents a given outcome of interest measured at the end of the evaluation year for teacher, administrator, or student j in school s in district d:

$$Y_{jsd} = \beta\, Treatment_{jsd} + \pi_d + \varepsilon_{jsd}. \quad (1)$$

We include fixed effects for randomization blocks, π_d. As these blocks are unique to district and school year, we do not include additional indicators for districts or years. We cluster our standard errors at the school level to account for the clustered experimental design. The coefficient, β, on the indicator for whether a teacher was in a school that was randomly assigned to treatment is our parameter of interest.

We designed the intervention to detect a treatment effect as small as 0.05 SD. However, it is possible that effects may be smaller, particularly for outcomes less proximal to the intervention (e.g., student test scores). Therefore, to increase statistical power when analyzing the effect of the intervention on students' test scores, we included a rich set of covariates. (For analyes of teacher-level outcomes, the only right-hand side variables included in our models are the treatment indicator and fixed effects for randomization block.) Student-level covariates include a cubic polynomial in prior-year same-subject test score, an interaction between student grade and prior-year same-subject test score, a linear term for prior-year opposite-subject test score with a dummy for those missing the opposite-subject test, grade-level indicators, gender, seven categories for race/ethnicity, an indicator for FRPL eligibility, an indicator for special education status, and an indicator for LEP students. Class- and school-by-grade-level covariates include the class-wide and school-by-grade-wide average of all student-level covariates, except that prior-year same-subject test score is only included linearly. In some of the student-level models, we also included teachers' prior-year value-added score in the same subject as the student test score. If the value-added score was missing, we imputed to the mean and included an indicator for missingness. All covariates were interacted with site and with subject. Impacts on student test scores shown below are similar when we exclude these covariates and only include fixed effects for randomization block.

We collected a range of outcome measures to allow us to identify the impact of the treatment on mediating outcomes, such as teacher and supervisor perceptions of the evaluation process, as well as student outcomes. However, statistical tests on multiple outcomes could lead us to observe a false positive due to multiple hypothesis testing. Therefore, within categories of outcomes that focus on similar underlying constructs (i.e., outcomes included in the same table), we adjust *p*-values with Bonferroni, Sidak, and Holm-Bonferroni corrections.

5. RESULTS

Relationship between Scores of Teacher-Selected and Unselected Videos

Giving teachers control of the cameras may have increased their willingness to use cameras, but allowing teachers to select which videos to submit for formal evaluation could have made it more difficult to identify teachers with poor instruction. In this section we compare how teachers performed on the videos submitted for formal evaluation with the other videos (up to eleven) recorded but not shared with a supervisor (but possibly including the videos submitted to the external content experts for nonevaluative feedback). An earlier study by Ho and Kane (2013) suggested the rankings of teaching practice using the videos teachers chose to share with their supervisors for high-stakes evaluations were similar to rankings on the full set of a teacher's videos. In Hillsborough County, Florida, teachers participating in the *Measures of Effective Teaching* project were allowed to choose which of their videos would be scored by their own principals. However, any of their videos could be scored by other principals and peer observers in

Hillsborough County. While the mean observation rating was 0.19 SD higher on the teacher-selected videos, the disattenuated correlation between a teacher's score on the videos submitted to supervisors and the remaining videos was approximately 1. While most teachers performed better on the selected videos (a signal that teachers understood the rubric, since they could identify which of their lessons would score better), the rankings were largely the same on the teacher-selected lessons as on the nonselected lessons.

In the present study, for each video chosen by a teacher to share with her administrator for her formal evaluation, we chose at random a video from the same period of the school year, which the teacher uploaded to our server but chose not to submit to her supervisor. We identified a sample of 197 such videos from a sample of sixty teachers randomly selected from the treatment group (thirty elementary, fifteen middle school math, fifteen middle school ELA). We contracted with a nonprofit organization, Teachstone, to score the videos using the CLASS observational rubric (Hamre, Pianta, and Choomat-Mooney 2009) and evaluated on four domains of teaching practice: Emotional Support, Classroom Organization, Instructional Support, and Student Engagement. (We did not reveal to Teachstone which videos had been submitted to a teacher's supervisor for formal evaluation.) Teachstone assigned eight raters to score middle school videos, and seven raters to score elementary videos. Each rater scored two videos—one chosen for high-stakes evaluation purposes and one not chosen for this purpose—from all thirty teachers in their grade range. Raters were certified on the CLASS rubric prior to the project and required to calibrate on four separate occasions during the project.

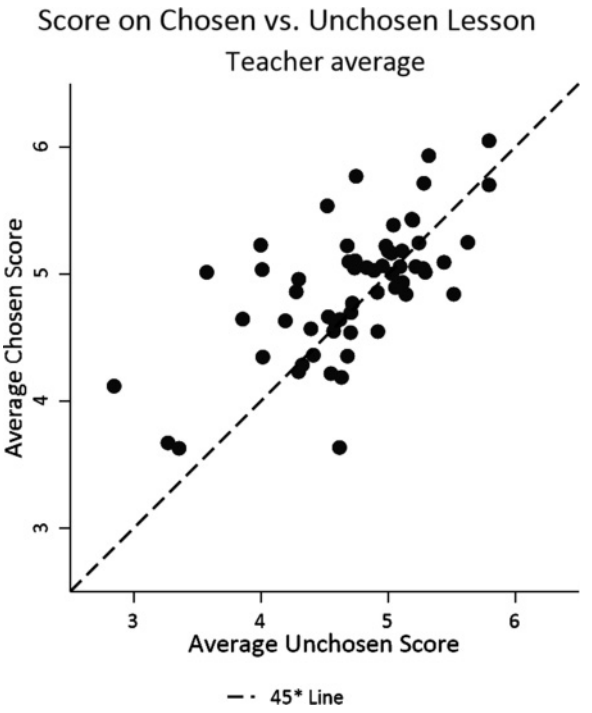
The mean scores on the videos chosen for formal, high-stakes evaluation were approximately 0.25 SD higher than the scores on teachers' other video. However, the disattenuated correlation between the two types of videos was moderately high (0.75).[8] In figure 1, we illustrate these patterns by presenting a scatterplot of lesson scores between the two sets of videos. We calculated the mean score for videos from each of these two groups, averaging over all the raters' scores. The horizontal axis measures the average score on the lessons that the teacher did not submit to his administrator, as scored by the observers in their grade grouping; the vertical axis measures the average score on the lessons that the same teacher chose to submit for formal evaluation. The dotted line in figure 1 represents the 45-degree line, along which scores would have been identical. For two-thirds of teachers, the average score on the lesson chosen for formal evaluation was higher than their other lessons. However, as shown in figure 1, the teachers who scored better on the lessons used for high-stakes evaluation also tended to score higher on the remaining videotaped lessons.[9] In other words, although teachers did

8. Following Ho and Kane (2013), we calculated the disattenuated correlation as follows:

$$\rho = \frac{Covariance(Score_{chosen,i,r}, Score_{unchosen,i,r'})}{\sqrt{rel_{chosen} * rel_{unchosen}}},$$

where $Score_{chosen,i,r}$ is the score of a video chosen for formal evaluation from teacher i by rater r, $Score_{unchosen,i,r'}$ is the score of an unchosen video from teacher i by a different rater r', and rel_{chosen} and $rel_{unchosen}$ are the reliability of chosen and unchosen video scores, respectively.

9. The measures used in figure 1 demonstrate a correlation of 0.64. The correlation differs from the disattenuated correlation reported earlier because it includes measurement error, as well as other variance components.



Notes: Each point is an average of seven scores each on chosen and unchosen videos for elementary teachers. For middle school teachers, each point is an average of eight scores each on chosen and unchosen videos.

Figure 1. Score on Videotaped Lessons Chosen for Formal Evaluation (i.e., "chosen" video on y-axis) versus other Videotaped Lessons Not Shared with Administrator (i.e., "unchosen" video on x-axis).

select their better videos to submit, the ranking of teachers' performance was largely the same as one would have gotten from watching any of the videos.

Impacts on the Evaluation Process

Table 4 reports the differences between the treatment and control groups on the number and types of observations reported by teachers and administrators on the end-of-year survey. Teachers in the treatment group reported that principals spent 1.58 hours less time in teachers' classrooms and completed 1.13 fewer in-person observations. The treatment principals did not recall spending significantly less time in treatment teachers' classrooms. However, they did report a net increase of 2.55 observations using video. Video-based observations in the control group were quite rare, with only 13 percent of control group principals reporting having done a video observation for one of the control group teachers. In other words, there was little evidence that the control group schools were implementing their own version of the treatment. Adjustments to *p*-values to account for multiple hypothesis testing (five tests conducted in table 4) yields the same pattern of results; *p*-values are larger in magnitude but those that were below 0.05 without the adjustment remain below the 0.05 threshold after adjustment.

Each week during the 2013–14 school year (cohort 1), we asked administrators in the treatment and control groups to describe the time devoted to various duties related to observations for a randomly selected teacher within the study sample. Table 5

Table 4. Impacts on the Number and Type of Observations

	Control Mean	Treatment − Control Difference (SE)	*N* (Teachers/ Administrators)
Teacher survey			
Teacher reported number of in-person observations supervisor did	4.42	−1.13*** (0.37)	392
Teacher reported hours supervisor spent doing in-person observations	5.24	−1.58*** (0.32)	389
Administrator survey			
Administrator reported average number of in-person observations	4.35	−0.04 (0.54)	115
Administrator reported average number of video observations	0.31	2.55*** (0.17)	95
Administrator reported doing any video observation	0.13	0.82*** (0.06)	95

Notes: The adjusted difference between control and treatment is the result of a regression of the dependent variable against fixed effects for randomization strata, a treatment indicator, and an indicator for whether the school is an elementary or middle school. Standard errors (SE) are reported in parentheses, and allow for clustering within school. A Bonferroni correction for five hypothesis tests changes the significance of the first result from the 99 percent level to the 95 percent level. Both the Sidak and Holm-Bonferroni corrections yield identical results.

*** Significant at the 99% level.

reports the results. In terms of the total time devoted to teacher observations, there was no difference between the treatment and control groups. Both groups spent slightly more than 41 minutes per week on various aspects of the observation process for a randomly selected teacher.[10] On average, the administrators in the treatment group spent 4.5 more minutes per week observing a randomly selected teacher than the control group. That is 45 percent more time observing than the control group mean of 10.1 minutes. Over the course of 20 weeks, that would amount to roughly 1.5 hours per teacher. However, the treatment group also reported spending less time on other aspects of the observation, such as completing forms. In an in-person observation, the observer needs to document what they saw, given the absence of a recording, and file the necessary paperwork. (This difference is no longer statistically significant when *p*-values are adjusted for multiple hypothesis testing.)

Although the intervention did not save time in the aggregate, administrators in the video group shifted their observation work to times of the day or week when classes were not being held and they could not have been performing in-person observations. We tracked the times when principals in each of the sites navigated into the observation viewing software. We compared the time stamps against the start and end of the school day and the scheduled lunch times at each school. We observed a total of 3,821 instances of principals navigating into the video viewing platform. Of these, roughly two-thirds (64 percent) of principal navigations occurred during noninstructional hours (before school, immediately after school, during lunch, in the evenings, on weekends, or holidays). This ranged from a low of 49 percent in Colorado to a high of 72 percent in

10. The average of 41 minutes per week includes 55 percent of surveys in which principals reported no observations for the randomly selected teacher identified in that week. Principals who indicated zero minutes spent observing that specific teacher may have spent time observing that teacher in other weeks, or observing other teachers that week.

Table 5. Impacts on Administrator Time Use

	Control Mean	Treatment − Control Difference (SE)	*N* (Administrators)
In minutes per week for a randomly chosen teacher			
Total	41.531	−0.119 (5.423)	105
Observing teachers	10.105	4.542*** (1.492)	105
Preparing to deliver feedback	4.5	0.434 (0.771)	105
Delivering feedback	5.617	−0.382 (0.715)	105
Pre-conference	2.445	−0.493 (0.450)	105
Scheduling an observation	2.029	−0.290 (0.370)	105
Writing the observation report	9.581	−1.592 (1.412)	105
Completing other forms for this teacher's observation	7.255	−2.338* (1.349)	105

Notes: The adjusted difference between control and treatment is the result of a regression of the dependent variable against fixed effects for randomization strata, a treatment indicator, and an indicator for whether the school is an elementary or middle school. Standard errors (SE) are reported in parentheses, and allow for clustering within school. Missing values on surveys that were otherwise completed were imputed as zero minutes. Surveys that were not returned were excluded. These measures only were available in the 2013–14 school year (cohort 1). A Bonferroni correction for eight hypothesis tests changes the significance such that the second result is significant at the 95 percent level, and the eighth result is no longer significant at the 90 percent level. Sidak and Holm-Bonferroni corrections yield identical results.

*Significant at the 90% level; ***significant at the 99% level.

Georgia. In our California district, nearly a quarter of administrator navigations (22 percent) occurred on weekends or holidays.

Impacts on Teacher and Administrator Impressions of the Evaluation Process

Table 6 compares the perceptions of treatment and control teachers of their school's evaluation processes at the end of the year, as well as after their first post-observation meeting. On the end-of-year survey, teachers in the treatment group were statistically significantly less likely to report that their conversations had been adversarial (5 percentage points less likely to say "almost always" or "often") or that they disagreed with the administrator about the appropriate score (7 percentage points less likely to say "almost always" or "often"). They were more likely to describe the observation process as "moderately fair" or "very fair" (10 percentage points). (Text in quotation marks comes directly from the survey.) Teachers in the treatment group were 14 percentage points more likely to identify a specific change in their practice resulting from post-observation conversations. They also were 9 percentage points more likely to report they had shared a video with a professional learning community or collaborative group at their school.

When surveyed soon after their first post-observation conference, teachers reported similar experiences to those they reported on the end-of-year survey. Treatment teachers were more likely than comparison teachers to report that administrators were "supportive," that administrators had "tried to take their perspective," that the conversations were "productive," and that they agreed with the administrator rating. Sample sizes are

Table 6. Impacts on Teacher Perceptions of the Observation Process

	Control Mean	Treatment − Control Difference (SE)	*N* (Teachers)
End of year survey			
Thinking about your post-conference . . .			
How well did observer understand your lesson plan and your goals for the class? (1 = "extremely" or "quite")	0.66	0.06 (0.05)	392
How often did you feel the conversation was adversarial? (1 = "almost always" or "often")	0.12	−0.05* (0.03)	389
How often did you and observer disagree about what actually happened during the lesson? (1 = "almost always" or "often")	0.05	−0.04** (0.02)	391
How often did you and observer disagree about the appropriate score for the lesson? (1 = "almost always" or "often")	0.08	−0.07*** (0.03)	390
Overall, how helpful was the feedback you received from your school administrator this year in helping you to improve your teaching? (1 = "extremely" or "quite")	0.37	0.15*** (0.05)	388
Overall, how fair was the classroom observation process this year? (1 = "very" or "moderate")	0.59	0.10** (0.04)	394
Can you identify a specific change in your teaching practice you made as a result of the feedback from your school administrator this year? (1 = "yes")	0.55	0.14*** (0.05)	390
Since January of this year, have you shared a video of your teaching in a professional learning community or other collaborative group? (1 = "yes")	0.10	0.09*** (0.03)	394
Post-Conference Survey			
How would you describe the relationship, at present, with your observer (e.g., principal or other instructional leader)? (1 = "very positive" or "somewhat positive")	0.83	0.06 (0.04)	308
How familiar is your observer with your strengths and weaknesses as a teacher? (1 = "very familiar" or "somewhat familiar")	0.71	0.05 (0.05)	309
To what extent did you agree or disagree with your observer or his/her recommendations? (1 = "completely agree" or "moderately agree")	0.65	0.13*** (0.05)	301
How supportive was your observer during the post-observation conference? (1 = "extremely" or "quite")	0.86	0.10*** (0.03)	300
During the post-observation conference, how much effort did your observer put into taking your perspective? (1 = "tremendous amount" or "quite a bit")	0.66	0.16*** (0.05)	298
How productive did you find the post-observation conference overall? (1 = "extremely" or "quite")	0.55	0.18*** (0.05)	300

Notes: For each comparison, the outcome is a dichotomous indicator for whether the respondent chose one of the top two categories on the Likert scale (e.g., on a 7-point Likert scale for agreement that ranges from "completely disagree" to "completely agree," the binary variable indicates a response choice of "completely agree" or "moderately agree"; Likert scales ranged across items from 5 points to 7 points). Results are similar when we keep the original scale. The adjusted difference between control and treatment is the result of a regression of the dependent variable against fixed effects for randomization strata, a treatment indicator, and an indicator for whether the school is an elementary or middle school. Standard errors (SE) are reported in parentheses, and allow for clustering within school. For teacher post-observation survey results we used only the first post-observation conference. We did not use all teacher responses, since treatment teachers were disproportionately likely to respond to all the post-observation surveys. A Bonferroni correction for eight hypothesis tests on the End of Year survey changes the significance such that the second, third, and sixth results are no longer significant at the 90 percent level. The fourth and eighth results remain significant at the 90 percent level only, while the fifth and seventh are significant at the 95 percent level. A Sidak correction yields identical results. A Holm-Bonferroni correction implies the fifth and seventh are significant at the 95 percent level while the fourth, sixth, and eighth are significant at the 90 percent level. A Bonferroni correction for six hypothesis tests on the Post-Conference surveys changes the significance such that the third and fifth post-conference results are significant at the 95 percent level, while the fourth and sixth results remain significant at the 99 percent level. Sidak and Holm-Bonferroni corrections yield identical results.

*Significant at the 90% level; **Significant at the 95% level; ***significant at the 99% level.

smaller in these analyses, as this survey was not administered to teachers in cohort 2.[11] After adjusting for multiple hypothesis testing, several results in table 6 are no longer statistically significant (see table notes). However, given that almost all of the

11. For teacher post-observation survey results we used only the first post-observation conference. We did not use all teacher responses, since treatment teachers were disproportionately more likely to respond to all the post-observation surveys.

Table 7. Impacts on Administrator Perceptions of the Observation Process

	Control Mean	Treatment − Control Difference (SE)	*N* (Administrators)
End of year survey			
Thinking about the teachers who were part of the study this year . . .			
How confident are you that your classroom observation provided an accurate rating of their teaching? (1 = "quite" or "extremely")	0.77	−0.06 (0.08)	117
Do you believe your post-conference meetings had a positive or negative impact on their subsequent instruction? (1 = "large positive" or "moderately positive")	0.67	0.03 (0.08)	116
How often were teachers defensive as you discussed your observation notes with them? (1 = "never" or "rarely")	0.66	0.23*** (0.07)	117
How often did you and the teacher disagree about what actually happened during the lesson? (1 = "never" or "rarely")	0.92	0.09** (0.04)	115
How often did you and the teacher disagree about the appropriate score for the lesson? (1 = "never" or "rarely")	0.92	0.06 (0.04)	115
How often did the following occur as a result of your classroom observations? (1 = "extremely often" or "quite often")			
I better understand my teachers' skills	0.89	−0.10 (0.07)	117
I better understood my teachers' development areas	0.77	0.02 (0.08)	117
I better understood what students were learning	0.87	−0.20** (0.08)	117
I gave teachers helpful feedback	0.64	0.07 (0.09)	117
I better understood the classroom challenges at my school	0.85	−0.23*** (0.08)	117
I helped my teachers reflect on their practice	0.70	0.04 (0.08)	117
Post-conference survey			
During the post-observation conference, how much effort did you put into taking teacher's perspective? (1 = "tremendous amount" or "quite a bit")	0.74	0.05 (0.09)	90
How would you describe your relationship at present? (1 = "very positive" or "somewhat positive")	0.96	−0.06 (0.06)	91
How productive did you find the post-observation conference overall? (1 = "extremely" or "quite")	0.84	−0.13 (0.09)	90

Notes: For each comparison, the outcome is a dichotomous indicator for whether the respondent chose one of the top two categories on the Likert scale (e.g., on a 4-point Likert scale for amount learned that ranges from "nothing" to "quite a bit," the binary variable indicates a response choice of "quite a bit" or "some"; Likert scales ranged across items from 4 points to 7 points). Results are similar when we keep the original scale. The adjusted difference between control and treatment is the result of a regression of the dependent variable against fixed effects for randomization strata, a treatment indicator, and an indicator for whether the school is an elementary or middle school. Standard errors (SE) are reported in parentheses, and allow for clustering within school. A Bonferroni correction for eleven hypothesis tests on the End-of-Year survey changes the significance such that the fourth and eighth results are no longer significant at the 90 percent level. The third and tenth results remain significant at the 95 percent level. Both the Sidak and Holm-Bonferroni corrections yield identical results.

**Significant at the 95% level; ***significant at the 99% level.

nonadjusted differences in table 6 are statistically significant, and the magnitude of these differences often are quite large, we conclude that the intervention had a strong positive effect on teachers' perception of the evaluation process.

Table 7 reports administrators' perceptions of the evaluation process. Regarding the adversarial nature of the post-observation discussions, responses of administrators were similar to those of teachers. Administrators in the treatment group were 23 percentage points more likely to report that teachers were "never" or "rarely" defensive during the post-observation conference. However, treatment administrators were not as confident as teachers that the video would lead to improvements in teachers' practice,

as treatment administrators were no more likely to report that post-conference meetings had a positive effect on their subsequent instruction. Administrators expressed some specific concerns about the video observations as a substitute for in-person observations. For instance, treatment administrators were at least 20 percentage points less likely to report they had a better understanding of student learning or classroom challenges in their school as a result of the classroom observation process. In other words, treatment administrators seemed to believe the video was a poor substitute for physical presence when it came to understanding students' learning.

Impacts on Assessments of Teachers' Instructional Practice

Table 8 reports impacts on teachers' self-assessment of their instructional practice and improvement. On the end-of-year survey, teachers were asked to rate their own instructional practice on several dimensions using a 5-point Likert scale. Teachers assigned to video-based observations rated their own practice lower than teachers in the comparison group. Treatment teachers were less likely to report themselves to be "quite proficient" or "extremely proficient" in terms of their ability to assess students' mastery of the content (10 percentage points), classroom management skills (8 percentage points), and ability to engage students in the curriculum (6 percentage points). These are all skills that would be observable in a video recording.

Yet, on these same dimensions, teachers in the video-based observation group were more likely to report their practice had improved during the year of the intervention. They reported their time management practices (8 percentage points) and lesson pacing (10 percentage points) had "improved somewhat more" or "improved much more" in the current year than in recent years. The treatment teachers were less likely to report their knowledge and understanding of their subject/field had improved during the year (10 percentage points). We cannot say whether this was due to the greater effectiveness of in-person observations in contributing to teachers' knowledge and understanding (although we believe this was unlikely), or whether the changes in time management and lesson pacing loomed larger in teachers' minds as a result of the video treatment. We are cautious in placing too much emphasis on the results presented in table 8, given that many of the differences are indistinguishable from zero when we account for multiple hypothesis testing.

For teachers in cohort 1, we also were able to examine effects on official observation scores, recorded by the supervisor in the evaluation system. These scores aim to capture teachers' overall effectiveness in the classroom, as assessed by school leaders on the evaluation rubrics provided by each district. Although the difference is positive (0.045 SD), it is not statistically distinguishable from zero.

Impacts on Teacher Retention

Table 9 reports impacts on teacher retention. This analysis excludes all cohort 2 teachers ($N = 88$), for whom we did not have administrative data in the year following the intervention. Additional teachers from cohort 1 were missing from these analyses because, in the district administrative data from the first year, they were not attached to students to allow us to examine the grade level that teachers taught ($N = 24$) or did not show up at all in course or staff files ($N = 6$).

Table 8. Impacts on Teacher Self-Assessments of Instructional Practice and Improvement

	Control Mean	Treatment − Control Difference (SE)	*N* (Teachers)
In thinking about your teaching practice, please rate your proficiency in the following areas . . . (1 = "extremely" or "quite" proficient)			
Assessing students' level of mastery of content/skills	0.87	−0.10** (0.04)	393
Using multiple methods of assessment of student learning	0.75	−0.02 (0.05)	393
Differentiating instruction for different learning styles	0.66	−0.02 (0.05)	392
Classroom management	0.84	−0.08** (0.03)	393
Engaging students in the curriculum	0.84	−0.06* (0.04)	393
How much did you learn this year about your practice in the following areas . . . (1 = "quite a bit" or "some")			
Time management practices	0.73	0.08** (0.04)	394
Lesson pacing	0.74	0.10** (0.04)	392
Your handling of student discipline and behavior	0.67	−0.04 (0.04)	394
Your knowledge and understanding of your main subject/field(s)	0.76	−0.10** (0.04)	393
The way that you make the material relevant	0.80	0.00 (0.04)	394
The rigor of your class assignments	0.82	−0.02 (0.04)	394
The way that you address specific students' social and emotional needs	0.72	−0.10** (0.04)	393
The way that you motivate your students	0.78	−0.03 (0.04)	392
The way you provide opportunities for student participation	0.84	0.04 (0.04)	394
The way you monitor student understanding	0.81	0.09** (0.04)	394

Notes: For each comparison, the outcome is a dichotomous indicator for whether the respondent chose one of the top two categories on the Likert scale (e.g., on a 4-point Likert scale for amount learned that ranges from "nothing" to "quite a bit," the binary variable indicates a response choice of "quite a bit" or "some"; Likert scales ranged across items from 4 points to 7 points). Results are similar when we keep the original scale. The adjusted difference between control and treatment is the result of a regression of the dependent variable against fixed effects for randomization strata, a treatment indicator, and an indicator for whether the school is an elementary or middle school. Standard errors (SE) are reported in parentheses, and allow for clustering within school. A Bonferroni correction for fifteen hypothesis tests changes the significance such that none of these results is significant at the 90 percent level. Both the Sidak and Holm-Bonferroni corrections yield identical results.

*Significant at the 90% level; **Significant at the 95% level.

Table 9. Impacts on Teacher Retention (Cohort 1)

	Control Mean	Treatment − Control Difference (SE)	*N* (Teachers)
Remain in same teaching assignment (school and grade)	0.558	0.099* (0.053)	321
Remain in same school	0.650	0.140*** (0.047)	321
Remain in same district	0.687	0.119*** (0.044)	321

Notes: The adjusted difference between control and treatment is the result of a regression of the dependent variable against fixed effects for randomization strata, and a treatment indicator. Standard errors (SE) are reported in parentheses and allow for clustering within school.

*Significant at the 90% level; ***significant 99% level.

Table 10. Impacts on Student Perceptions of the Classroom Environment

	Pooled Year 1				Cohort 1 Follow-Up Year			
	Control Mean	Treatment − Control Difference (SE)	*N* (Students)	*N* (Teachers)	Control Mean	Treatment − Control Difference (SE)	*N* (Students)	*N* (Teachers)
Average of all items	−0.004	0.020 (0.030)	12,862	389	0.008	0.035 (0.042)	3,696	113
Classroom management construct	−0.025	0.037 (0.034)	12,860	389	0.050	0.076 (0.046)	3,695	113

Notes: The adjusted difference between control and treatment is the result of a regression of the dependent variable against fixed effects for randomization strata, a treatment indicator, and an indicator for whether the school is an elementary or middle school. Standard errors (SE) are reported in parentheses and allow for clustering within school. Results are similar when controlled for observable student background characteristics; however, the sample size is reduced by roughly one third due to challenge merging student identifiers from survey rosters to those in administrative data.

We find that the intervention had statistically significant impacts on all three types of teacher retention: (1) whether teachers remained in their same teaching assignment (same school and grade), (2) remained in their same school, or (3) remained in the same district. Compared with 56 percent of control group teachers, 66 percent of treatment teachers maintained their teaching assignment in the year following the intervention. Compared with 71 percent of control group teachers, 84 percent of treatment teachers remained in their school. Finally, compared with 78 percent of control group teachers, 89 percent of treatment teachers remained in their same district or the teaching profession. All three of these differences are statistically significant.

Differences in retention rates between the treatment and control group are quite large relative to previous studies (Borman and Dowling 2008; Papay et al. 2017). Prior research also has suggested that teachers' relationships with principals is a primary factor driving attrition (Boyd et al. 2011; Kraft, Marinell, and Yee 2016). One likely explanation is that the impact on attrition was mediated by more supportive, less confrontational relationships between teachers and administrators.[12]

Impacts on Student Outcomes

For both sets of student outcomes—the student perception survey and student test scores—we collected data at the end of the intervention year and, for the first cohort of teachers, at the end of the following year. These data allowed us to test for any lagged impacts on student outcomes as teachers potentially incorporated changes in their practice in the subsequent year.

Table 10 reports effects of the intervention on students' survey responses. We find no effect of the program on a composite measure or on a subset of items focused on classroom management, either at the end of the intervention year or in the follow-up year. Our estimates are all positively signed, and magnitudes for effects on the classroom management construct (0.037 SD at the end of the intervention year, and 0.076 SD in the follow-up year) are on par with other educational interventions (Fryer 2017).

12. In Appendix table A.2, we investigated whether the effects on retention differed for teachers who had strong prior relationships with principals or for those who had strong baseline observation ratings or for those with more teaching experience. None of those interactions was statistically different from zero.

Table 11. Impacts on Student Test Scores

	Pooled Year 1			Cohort 1 Follow-Up Year		
	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
Combined math and English language arts						
Treatment	−0.042	0.003	−0.009	−0.058	0.031	0.015
	(0.071)	(0.025)	(0.022)	(0.098)	(0.033)	(0.033)
Adjusted R^2	0.004	0.638	0.644	0.008	0.684	0.687
N (Teachers)	339	339	339	194	194	194
N (Students)	20,575	20,575	20,575	11,834	11,834	11,834
Math only						
Treatment	−0.001	0.018	−0.031	−0.041	0.049	0.008
	(0.077)	(0.035)	−0.029	(0.109)	(0.042)	(0.039)
Adjusted R^2	0.005	0.652	0.660	0.011	0.704	0.708
N (Teachers)	222	222	222	122	122	122
N (Students)	10,504	10,504	10,504	5,928	5,928	5,928
English language arts only						
Treatment	−0.069	−0.015	−0.032	−0.061	0.017	0.013
	(0.073)	(0.024)	(0.024)	(0.103)	(0.033)	(0.034)
Adjusted R^2	0.005	0.631	0.635	0.007	0.675	0.675
N (Teachers)	232	232	232	124	124	124
N (Students)	10,071	10,071	10,071	5,906	5,906	5,906

Notes: The adjusted difference between control and treatment is the result of a regression of the dependent variable against fixed effects for randomization strata, and a treatment indicator. Models 2-3 also include student- and class-level covariates, and Model 3 includes prior teacher value added as a covariate. Student-level covariates include a cubic polynomial in prior-year same-subject test score, an interaction between student grade and prior-year same-subject test score, a linear term for prior-year opposite-subject test score with a dummy for those missing the opposite-subject test, grade-level indicators, gender, seven categories for race/ethnicity, an indicator for free or reduced price lunch eligibility, an indicator for special education status, and an indicator for students with Limited English Proficiency. Class-level covariates include the class-wide average of all student-level covariates, except that prior-year same-subject test score is only included linearly. Teacher value added is the teacher's same-subject value added for the 2012–13 and 2013–14 school years, with a dummy for those missing value added. All covariates were interacted with site and additionally with subject for the combined model. Samples include nontreatment students and teachers in order to increase the precision of estimates for these covariates, and models include an indicator for this group. After controlling for student background characteristics, nontreatment students generally do not perform differently than treatment students on end-of-year tests.

However, standard errors are large relative to the point estimates. It is possible that there were small effects on student perceptions of teacher behaviors, but we were underpowered to detect effects of such magnitude.[13]

Table 11 reports effects on students' test scores. As with student survey outcomes, we do not find any impact on student achievement. This is true when pooling and disaggregating by subject areas (math and ELA), and across school years. Here, we are able to control for a range of student and class characteristics in an effort to increase precision. Compared with Model 1, which includes no covariates, Model 2 includes student- and class-level covariates, and Model 3 adds prior teacher value-added. We also include nonparticipating students and teachers in order to increase the precision of estimates for

13. We aimed to increase the precision of our estimates by controlling for observable background characteristics of students. However, we lose roughly one-third of the sample because roster IDs used to identify survey responses don't match administrative data IDs where we are able to collect demographic information. Results are similar in the remaining, smaller sample.

these covariates (along with an indicator for this group). However, by and large, effects are close to 0 SD.

6. CONCLUSION

Given the longstanding evidence of variability in individual teacher effects on student achievement, teacher evaluation will remain a tempting lever for policy makers seeking to improve student outcomes. For teachers without tenure protection (e.g., during the initial probationary period), such evaluations could prove useful in selecting out ineffective teachers, even if they were not useful in improving a given teacher's performance. However, for the vast majority of teachers who enjoy the job protections afforded to nonprobationary teachers, the value of the teacher evaluation process depends heavily on demonstrated improvement in teaching behavior. Those improvements would have to be large enough to justify the considerable investments of teachers' and principals' time devoted to classroom observations.

Our evidence suggests that, among a sample of volunteers, allowing teachers to submit video in lieu of in-person observations does improve the evaluation process in several ways. Both teachers and administrators reported that post-observation discussions were less adversarial. After witnessing their own teaching on video, teachers were more self-critical, especially with respect to time management and questioning. Teachers also were more likely to identify a specific change they made in their practice as a result of observation and feedback. In addition, principals reallocated their observation duties to noninstructional hours, when the opportunity cost (or shadow price) of their time was lower.[14] Because administrators in the treatment group indicated they had lower levels of understanding of student learning, districts or schools that implement similar programs may consider maintaining at least one in-person observation.

Perhaps reflecting the impacts on teacher–supervisor relationships, the intervention also had large positive impacts on teacher retention. Treatment teachers were more likely than control group participants to remain in their same teaching assignment—in the same school and grade—in the following year. They also were more likely to remain in their school, their district, or the teaching profession. Effect sizes in the range of 9 to 12 percentage points are substantively larger than those from studies examining other factors influencing teacher retention (for one meta-analysis, see Borman and Dowling 2008). Given these positive impacts, our team created an online toolkit that school and district leaders may use when considering implementing (or adapting) a similar intervention to the one described here.[15]

However, in our study, we did not find impacts on students' reports of their classroom experiences or performance on end-of-year state tests. We hypothesized two possible costs to using videotaped lessons for formal observation: First, teachers' self-selection of lessons to share with administrators may make the process less informative. Even if the rankings of teachers were unchanged, teachers may have been able to conceal their specific instructional weaknesses from their supervisors. Second, the intervention may have crowded out teacher preparation for other nonrecorded lessons.

14. We cannot say just how valuable the time reallocation was to principals. However, because principals could have done their video observations during instructional hours and chose not to do so, we assume this was beneficial.
15. See https://cepr.harvard.edu/video-observation-toolkit.

Another possible explanation is that the impacts on teacher instruction may have required more time to emerge. Yet, prior research on teacher coaching indicates that treatment effects generally show up both on teachers' instructional practice and student test scores during the intervention year (Kraft, Blazar, and Hogan 2018; for an exception, see Allen et al. 2011). Taylor and Tyler (2012) also found impacts of teacher evaluation, both in the intervention year and in subsequent years. Therefore, we hypothesize that tracking the effect of our program for additional years is unlikely to show positive effects on student perceptions of the classroom environment or student test scores.

We interpret our finding of positive effects on teacher and principal perceptions of the evaluation process and the reduction in teacher retention, and null findings on student outcomes, is most likely explained by the fact that the intervention was not strong enough to generate instructional change. Following from the literature on the success of teacher coaching programs, we hypothesize that if the use of video in classroom observations and evaluation processes is to generate improvements in student outcomes—and not simply improve teachers' and principals' perception of the process—video feedback may need to be paired with specific instructional suggestions that teachers could practice and resubmit to their supervisors.

ACKNOWLEDGMENTS

We gratefully acknowledge generous support from the Ken and Anne Griffin Foundation and the Bill & Melinda Gates Foundation. Min Lee, from the Ken and Anne Griffin Foundation, and Steve Cantrell, from the Bill & Melinda Gates Foundation, provided many helpful insights throughout the project. We also thank our partners employed in state departments of education and school districts, and our collaborators from technology and video-coaching vendors, The New Teacher Project, thereNow, BloomBoard, and Swivl. Without their help, this work would not have been possible. Most importantly, we are grateful for the participation of the principals, assistant principals, and teachers—the time and perspectives they shared have informed and pushed our thinking throughout the project. This paper uses a combination of project-collected and district administrative data. The project-collected survey data can be obtained from the authors beginning six months after publication through three years hence. The district administrative data can be obtained by filing a request directly with each of the participating districts; the authors are willing to provide guidance on how it may be acquired.

REFERENCES

Allen, Joseph P., Robert C. Pianta, Anne Gregory, Amori Yee Mikami, and Janetta Lun. 2011. An interaction-based approach to enhancing secondary school instruction and student achievement. *Science* 333(6045): 1034–1037.

Araujo, Maria, Pedro Carneiro, Yyannú Cruz-Aguayo, and Norbert Schady. 2014. A helping hand? Teacher quality and learning outcomes in kindergarten. Washington, DC: Inter-American Development Bank Working Paper Series No. IDB-WP-665.

Atteberry, Allison, Susanna Loeb, and James Wyckoff, J. 2017. Teacher churning: Reassignment rates and implications for student achievement. *Educational Evaluation and Policy Analysis* 39(1): 3–30.

Ballou, Dale, and Matthew G. Springer. 2015. Using student test scores to measure teacher performance: Some problems in the design and implementation of evaluation systems. *Educational Researcher* 44(2): 77–86 .

Bell, Courtney, Drew Gitomer, Daniel McCaffrey, Bridget Hamre, Robert Pianta, and Yi Qi. 2012. An argument approach to observation protocol validity. *Educational Assessment* 17(2–3): 62–87.

Blazar, David. 2015a. Effective teaching in elementary mathematics: Identifying classroom practices that support student achievement. *Economics of Education Review* 48:16–29.

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

Borman, Geoffrey D., and N. Maritza Dowling. 2008. Teacher attrition and retention: A meta-analytic and narrative review of the research. *Review of Educational Research* 78(3): 367–409.

Boyd, Donald, Pam Grossman, Marsha Ing, Hamilton Lankford, Susanna Loeb, and James Wyckoff. 2011. The influence of school administrators on teacher retention decisions. *American Educational Research Journal* 48(2): 303–333.

Brunvand, Stein, and Barry Fishman. 2006. Investigating the impact of the availability of scaffolds on preservice teacher noticing and learning from video. *Journal of Educational Technology Systems* 35(2006): 151–174.

Danielson, Charlotte. 2011. *Enhancing professional practice: A framework for teaching*. Alexandria, VA: Association for Supervision and Curriculum Development.

Dee, Thomas, and James Wyckoff. 2015. Incentives, selection, and teacher performance: Evidence from IMPACT. *Journal of Policy Analysis and Management* 34(2): 267–297.

Dynarski, Mark. 2016. *Teacher observations have been a waste of time and money*. Available https://www.brookings.edu/research/teacher-observations-have-been-a-waste-of-time-and-money/. Accessed 19 February 2020.

Fryer , Roland. 2017. The production of human capital in developed countries: Evidence from 196 randomized field experiments. In *Handbook of economic field experiments*, Vol. 2, edited by Abhijit Vinayak Banerjee and Esther Duflo, pp. 95–322. Amsterdam: North-Holland.

Fulbeck, Eleanor, Martyna Citkowicz, Candace Hester, David Manzeske, Melissa Yisak, and Ryan Eisner. 2016. Newark Public Schools and Newark Teachers Union teacher contract evaluation: Year 1 report. Washington, DC: American Institutes for Research.

Grissom, Jason A., Susanna Loeb, and Benjamin Master. 2013. Effective instructional time use for school leaders: Longitudinal evidence from observations of principals. *Educational Researcher* 42(8): 433–444.

Hamre, Bridget K., Robert C. Pianta, and Lia Chomat-Mooney. 2009. Conducting classroom observations in school-based research. In *Conducting science-based psychology research in schools*, edited by Lisa M. Dinella, pp. 79–105. Washington, DC: APA Press.

Hanushek, Eric A., and Steven G. Rivkin. 2010. Generalizations about using value-added measures of teacher quality. *American Economic Review* 100(2): 267–271.

Hayton, James C., David G. Allen, and Vida Scarpello. 2004. Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational Research Methods* 7(2): 191–205.

Hill, Heather C., Charalambos Y. Charalambous, and Matthew A. Kraft. 2012. When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher* 41(2): 56–64.

Hill, Heather, and Pam Grossman. 2013. Learning from teacher observations: Challenges and opportunities posed by new teacher evaluation systems. *Harvard Educational Review* 83(2): 371–384.

Ho, Andrew D., and Thomas J. Kane. 2013. The reliability of classroom observations by school personnel. Seattle, WA: Bill & Melinda Gates Foundation.

Jiang, Jennie Y, Susan Sporte, and Stuart Luppescu. 2015. Teacher perspectives on evaluation reform: Chicago's REACH students. *Educational Researcher* 44(2): 105–116.

Kane, Thomas J., and Steve Cantrell. 2010. Learning about teaching: Initial findings from the Measures of Effective Teaching project. Seattle, WA: Bill & Melinda Gates Foundation.

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

Kane, Thomas J., Daniel F. McCaffrey, Trey Miller, T., and Douglas O. Staiger. 2013. Have we identified effective teachers? Validating measures of effective teaching using random assignment. Seattle, WA: Bill & Melinda Gates Foundation.

Kleinknecht, Marc, and Jürgen Schneider. 2013. What do teachers think and feel when analyzing videos of themselves and other teachers teaching? *Teaching and Teacher Education* 33:13–23.

Kline, Paul. 1994. *An easy guide to factor analysis*. London: Routledge.

Kraft, Matthew A., David Blazar, and Dylan Hogan. 2018. The effect of teacher coaching on instruction and achievement: A meta-analysis of the causal evidence. *Review of Educational Research* 8(4): 547–588.

Kraft, Matthew A., Eric J. Brunner, Shaun M. Dougherty, and David J. Schwegman. 2019. Teacher evaluation reform and the supply and quality of new teachers. Unpublished paper, Brown University.

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

Kraft, Matthew A., Will Marinell, and Derek Shen-Wei Yee. 2016. School organizational contexts, teacher turnover, and student achievement: Evidence from panel data. *American Educational Research Journal* 53(5): 1411–1449.

Ost, Benjamin. 2014. How do teachers improve? The relative importance of specific and general human capital. *American Economic Journal. Applied Economics* 6(2): 127–151.

Papay, John P., Andrew Bacher-Hicks, Lindsay C. Page, and Will H. Marinell. 2017. The challenge of teacher retention in urban schools: Evidence of variation from a cross-site analysis. *Educational Researcher* 46(8): 434–448.

Raymond, Diana D., Peter W. Dowrick, and Chris L. Kleinke. 1993. Affective responses to seeing oneself for the first time on unedited videotape. *Counselling Psychology Quarterly* 6(3): 193–200.

Rosaen, Cheryl L., Mary Lundeberg, Marjorie Cooper, Anny Fritzen, and Marjorie Terpstra. 2008. Noticing noticing: How does investigation of video records change how teachers reflect on their experiences? *Journal of Teacher Education* 59(4): 347–360.

Santagata, Rossella, and Giulia Angelici. 2010. Studying the impact of the lesson analysis framework on preservice teachers' abilities to reflect on videos of classroom teaching. *Journal of Teacher Education* 61(4): 339–349.

Sartain, Lauren, and Matthew P. Steinberg. 2016. Teachers' labor market responses to performance evaluation reform: Experimental evidence from Chicago Public Schools. *Journal of Human Resources* 51(3): 615–655.

Sawchuk, Stephen. 2016. *ESSA loosens reins on teacher evaluations, qualifications*. Available https://www.edweek.org/ew/articles/2016/01/06/essa-loosens-reins-on-teacher-evaluations-qualifications.html. Accessed 24 April 2019.

Stecher, Brian M., Deborah J. Holtzman, Michael S. Garet, Laura S. Hamilton, John Engberg, Elizabeth D. Steiner, Abby Robyn, Matthew D. Baird, Italo A. Gutierrez, Evan D. Peet, Iliana Brodziak de los Reyes, Kaitlin Fronberg, Gabriel Weinberger, Gerald P. Hunter, and Jay Chambers. 2019. Intensive partnerships for effective teaching enhanced how teachers are evaluated but had little effect on student outcomes. Santa Monica, CA: RAND Corporation.

Steinberg, Matthew P., and Lauren Sartain. 2015. Does teacher evaluation improve school performance? Experimental evidence from Chicago's excellence in teaching project. *Education Finance and Policy* 10(4): 535–572.

Taylor, Eric S., and John H. Tyler. 2012. The effect of evaluation on teacher performance. *The American Economic Review* 102(7): 3628–3651.

U.S. Department of Education (USDOE). 2012. *Public school principal data file*. Available https://nces.ed.gov/surveys/sass/tables_list.asp. Accessed 18 February 2020.

Young, Alwyn. 2018. Channeling Fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results. *The Quarterly Journal of Economics* 134(20): 557–598.

APPENDIX A

Table A.1. External Validity: Study Participants versus Nonparticipants across Districts

	Colorado		Delaware	
	Study Participants	Participating – Nonparticipating difference (SE)	Study Participants	Participating – Nonparticipating difference (SE)
District characteristics				
Proportion FRPL-eligible	0.412	−0.065 (0.074)	0.497	−0.007 (0.108)
Proportion IEP	0.000	0.000 (0.000)	0.122	−0.002 (0.041)
Proportion LEP	0.070	0.001 (0.036)	0.034	−0.001 (0.026)
Proportion African American	0.006	−0.007 (0.011)	0.364	−0.035 (0.120)
Proportion Asian	0.007	−0.003 (0.006)	0.042	0.013 (0.019)
Proportion Hispanic	0.240	−0.033 (0.089)	0.109	0.013 (0.039)
Proportion Native American	0.062	0.054*** (0.011)	0.004	−0.001 (0.001)
Proportion Pacific Islander	0.001	−0.001 (0.002)	0.002	0.000 (0.001)

Table A.1. Continued.

	Colorado		Delaware	
	Study Participants	Participating – Nonparticipating difference (SE)	Study Participants	Participating – Nonparticipating difference (SE)
Proportion white	0.657	−0.017 (0.092)	0.456	0.003 (0.106)
Proportion multiple/other race	0.026	0.007 (0.007)	0.023	0.008 (0.007)
Total enrollment	4,991	176 (5,213)	5,200	2,481 (1,674)
Total Teacher FTE	289	18 (286)	370	179 (116)
Student–Teacher ratio	15.762	1.860 (1.685)	13.639	0.419 (1.321)
Share charter	0.000	−0.006 (0.031)	0.500	−0.083 (0.198)
Share urban/suburban	0.167	−0.013 (0.160)	0.500	−0.306* (0.167)

Notes: The district data are drawn from the U.S. Department of Education National Center for Education Statistics Common Core of Data, from the year prior to our study (school year 2012–13). Only local school districts and charter agencies are included; supervisory union administrative centers, regional education service agencies, state agencies, and federal agencies are excluded. Means and differences are computed giving each district equal weight. Differences are calculated from a linear regression controlling for treatment status and no other covariates. Standard errors (SE) are reported in parentheses. FRPL = free or reduced-price lunch; IEP = individualized education plan; LEP = limited English proficiency; FTE = full-time equivalent.

***Significant at the 99% level.

Table A.2. Heterogeneous Impacts on Teacher Retention (Cohort 1)

	Main Effect	Interaction Effect	*N* (Teachers)
Good prior relationship with administration			
Remain in same teaching assignment (school and grade)	0.042 (0.094)	0.126 (0.110)	299
Remain in same school	0.073 (0.088)	0.133 (0.099)	299
Remain in same district	0.035 (0.091)	0.153 (0.106)	299
High administration baseline rating			
Remain in same teaching assignment (school and grade)	0.002 (0.080)	0.263*** (0.094)	259
Remain in same school	0.101 (0.073)	0.085 (0.088)	259
Remain in same district	0.074 (0.074)	0.066 (0.087)	259
Ten or more years of experience			
Remain in same teaching assignment (school and grade)	0.143 (0.090)	−0.051 (0.103)	300
Remain in same school	0.171** (0.075)	−0.021 (0.085)	300
Remain in same district	0.151** (0.074)	−0.028 (0.080)	300

Table A.2. Continued.

Notes: Our model regresses the dependent variable against fixed effects for randomization strata, an indicator for our mediator of interest (good prior relationship/high baseline rating/ten or more years of experience), a treatment indicator, and the interaction between our mediator and the treatment indicator. Standard errors are reported in parentheses, and allow for clustering within-school. Good prior relationship with administration is defined as answering "Very" or "Quite" to the question "Overall, how much do you enjoy working with your administrator" on the baseline survey. Seventy-five percent of teachers fell into this category. High administration baseline rating is defined as a teacher's administrator answering "top 5%" or "top 25%" to the question "Among all the teachers you have known who taught the same grade/subject, how would you rate the overall quality of instruction provided by this teacher?" on the baseline survey. Fifty-one percent of teachers fell into this category. Years of experience are taken from our baseline survey. Sixty-two percent of teachers reported having ten or more years of experience.

**Significant at the 95% level; ***significant at the 99% level.