ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 19-62

Can Teacher Evaluation Systems Produce High-Quality Feedback? An Administrator Training Field Experiment

Matthew A. Kraft
Brown University

Alvin Christian
University of Michigan

A core motivation for the widespread teacher evaluation reforms of the last decade was the belief that these new systems would promote teacher development through high-quality feedback. We examine this theory by studying teachers' perceptions of evaluation feedback in Boston Public Schools and evaluating the district's efforts to improve feedback through an administrator training program. Teachers generally reported that evaluators were trustworthy, fair, and accurate, but that they struggled to provide high-quality feedback. We find little evidence the training program improved perceived feedback quality, classroom instruction, teacher self-efficacy, or student achievement. Our results illustrate the challenges of using evaluation systems as engines for professional growth when administrators lack the time and skill necessary to provide frequent, high-quality feedback.

VERSION: June 2021

Suggested citation: Kraft, Matthew A., and Alvin Christian. (2021). Can Teacher Evaluation Systems Produce High-Quality Feedback? An Administrator Training Field Experiment. (EdWorkingPaper: 19-62). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/ydke-mt05

Can Teacher Evaluation Systems Produce High-Quality Feedback? An Administrator Training Field Experiment

Matthew A. Kraft
Brown University

Alvin Christian
University of Michigan

May 2021

Abstract

A core motivation for the widespread teacher evaluation reforms of the last decade was the belief that these new systems would promote teacher development through high-quality feedback. We examine this theory by studying teachers' perceptions of evaluation feedback in Boston Public Schools and evaluating the district's efforts to improve feedback through an administrator training program. Teachers generally reported that evaluators were trustworthy, fair, and accurate, but that they struggled to provide high-quality feedback. We find little evidence the training program improved perceived feedback quality, classroom instruction, teacher self-efficacy, or student achievement. Our results illustrate the challenges of using evaluation systems as engines for professional growth when administrators lack the time and skill necessary to provide frequent, high-quality feedback.

Keywords: evaluation feedback quality, evaluator training, observation and feedback, randomized controlled trial, teacher evaluation

We are extremely grateful to Jonathan Barrows, Jerome Doherty, Jared Joiner, Jessica Madden-Fuoco Emily Qazilbash, Ross Wilson and many other members of the Boston Public Schools for their support of this research. All errors and omissions are our own.

Can Teacher Evaluation Systems Produce High-Quality Feedback? An Administrator Training Field Experiment

Over the last decade, nearly every state in the U.S. has implemented major reforms to their teacher evaluation systems (Donaldson & Papay, 2015; Steinberg & Donaldson, 2016). A twofold theory of action motivated these reforms: differentiating teacher performance for accountability (Hanushek, 2009; Thomas et al., 2010) and promoting professional development through classroom observations and feedback (Almy, 2011; Curtis & Wiener, 2012; Papay, 2012). On paper, most states have emphasized the latter goal of using evaluation to improve teachers' instruction (Center on Great Teachers and Leaders, 2014). In practice, few new systems have provided the necessary training or support to develop administrators' capacity to deliver high-quality evaluation feedback.

A growing number of scholars have documented administrators' limited efficacy in providing evaluation feedback, stressing the importance of developing training programs to build their feedback skills (Donaldson, 2012; Feeney, 2007; Sartain et al., 2011). Administrators themselves have expressed strong interest in receiving professional development on coaching teachers (Kraft & Gilmour, 2016; Sporte et al., 2013). However, what little training administrators have received as part of new evaluation systems has largely been focused on improving their reliability as classroom observers for accountability purposes (Herlihy et al., 2014; Bell et al., 2016). We still know very little about how to improve the quality of feedback administrators provide.

In this paper, we study one district's efforts to better support administrators' capacity to promote teacher development through the evaluation process. Boston Public Schools (BPS) implemented major reforms to its teacher evaluation system in 2011-12 with a focus on using evaluation as a tool for professional growth. The following year, BPS convened a group of

experienced administrators to adapt, pilot, and refine an evaluator feedback training program developed by the New Teacher Center. District administrators then participated in the semester-long training program in small cohorts staggered across the 2013-14 and 2014-15 school years.

We partnered with BPS to examine the implementation of the new evaluation system, estimate the effect of the evaluator training program, and explore how the district might further strengthen the evaluation feedback teachers receive. We ask three research questions:

> *What are teachers' perceptions of the new evaluation system in BPS and the quality of feedback they receive?*
>
> *Can training administrators to provide high-quality evaluation feedback improve teachers' perceptions of feedback quality, classroom instruction, teacher self-efficacy, and student achievement?*
>
> *What predicts teachers' perceptions about their evaluation feedback quality?*

We answer the first question using teachers' responses to a district-wide survey we developed and administered to capture teachers' experiences with the evaluation system and perceptions about the feedback they received. We answer the second question by exploiting the staggered rollout of the training program and randomly assigning school-based evaluation teams to attend in one of four semesters. We explore the third question by combining responses from our teacher survey with district administrative data that links administrators with the teachers they evaluated.

Our findings reveal both the potential and limitations of promoting professional development through the teacher evaluation process. Throughout our analyses we focus on teachers' *perceptions* of feedback quality because both pedagogical theory and prior empirical research suggest that teachers are unlikely to respond to feedback in constructive ways if they do not believe it to be accurate or useful (Garubo & Rothstein, 1998; Feeney, 2007; Cherasaro et al., 2016). We find that while teachers reported being observed frequently and receiving regular feedback, only one out of four felt that evaluation feedback helped them improve their practice.

Results from our randomized field trial suggest the intensive training program had little impact on administrators' feedback skills or the subsequent outcomes we hypothesized it might affect. We find relatively precise null effects on the frequency and length of post-observation meetings, teachers' perceptions of evaluation feedback quality, and student achievement. Further analyses suggest that these null effects are not driven by low program quality or a lack of take-up. However, our exploratory analyses reveal that it is possible for administrators to provide high-quality feedback within the evaluation process. We find that some administrators are perceived to be far more effective at providing feedback than others. Administrators' experience level and the racial match between teachers and administrators appear to play important roles in shaping perceived feedback quality.

This study makes several important contributions to research, policy, and practice. Our findings advance our understanding of how the design and implementation of evaluation systems shape feedback quality. We provide among the first rigorous experimental evidence on the efficacy of efforts to improve the feedback administrators provide within the high-stakes evaluation process. We also build on prior studies that examine teachers' experiences with high-stakes evaluation systems by collecting rich descriptive data on the nature of the evaluation feedback administrators provide. Finally, we quantify for the first time the large variation across administrators in their ability to provide effective feedback. Together, these findings can inform states' and districts' ongoing efforts to redesign teacher evaluation systems under the increased flexibility provided by the Every Student Succeeds Act (ESSA).

Theory and Prior Literature

The Theory of Action Behind Performance Feedback

A wide body of literature in organizational psychology, management sciences, and education research has focused on the practice of providing performance feedback. Fundamentally, feedback is information communicated to a person that is intended to modify his or her thinking or behavior to improve task performance (Shute, 2008). A seminal meta-analysis of feedback interventions by Kluger and DeNisi (1996) documents both large positive effects, on average, and wide heterogeneity where a sizable fraction of interventions decreased task performance. They explain this phenomenon based on the moderating effect of how feedback shapes learners' locus of attention. Feedback that is about a specific task generally promotes learning, while feedback that focuses on the learner can impede performance.

The broader literature also identifies a range of factors that influence whether feedback improves performance including the nature of the feedback (timing, frequency, and accuracy), the context (trust in the evaluator and the perceived fairness of the process), and the orientation of the person receiving feedback (openness to feedback, perceived need for change, belief that change is feasible, and a willingness to take action) (DeNisi & Sonesh, 2011; Smither et al., 2005). Education research on the use of feedback to promote teacher improvement suggests that feedback is most effective when it is immediate, specific, and actively engages teachers (Scheeler et al., 2004; Thurlings et al., 2013).

In the last decade, teacher evaluation reforms have made performance feedback a regular feature of the U.S. public education system (Grissom & Youngs, 2016; Tuma et al., 2018). We draw on existing theory and evidence to develop a theory of action for how performance feedback might promote instructional improvement as part of the teacher evaluation process. As shown in Figure 1, the theory begins with a trusted and reputable evaluator observing a teacher's instruction using a set of clearly defined criteria (McLaughlin & Pfeifer, 1988; Kimball, 2002).

Second, the evaluator and the teacher meet to discuss the observation and collectively work to identify the teacher's individual learning goals. At the core of these conversations are the descriptive data and objective observations collected during class observations.

The actual act of providing feedback during the post-observation meeting can take multiple forms and might vary based on evaluators' expertise, teachers' experience, and teachers' performance (Danielson & McGreal, 2000). Evaluators might make targeted recommendations for instructional resources and professional development programs to help teachers meet their learning goals. Evaluators might engage in reflective coaching by asking teachers to analyze their own instruction (Glickman, 2002). Finally, evaluators might provide more directive feedback on how teachers can take specific steps to improve their instruction. These different techniques can be used both in sequence over multiple feedback cycles or in combination during a single post-observation meeting.

Teachers must then take action to translate this feedback into changes in their instructional practice. Teachers might engage with the recommended resources and professional development opportunities; they might self-direct their own learning; or they might modify their instruction based on the specific feedback they received. The theory of action then posits that these actions will lead to improvements in teachers' instructional practice. The feedback cycle then repeats with evaluators and teachers assessing the progress made towards the learning goals, continuing to refine practices, and trying new approaches. Finally, these improvements in instructional quality will lead to increases in two mutually self-reinforcing outcomes: teacher self-efficacy and student achievement (Johnson & Birkeland, 2003; Zee & Koomen, 2016).

Research suggests that there are several necessary conditions for the theory of action described in Figure 1 to be successful. Evaluators must have the capacity to observe and meet

with teachers regularly, develop trusting relationships with their teachers, and be trained to provide clear, specific, and actionable feedback (Garet et al., 2001; Desimone, 2009; Desimone & Garet, 2015). Relational trust between administrators and teachers plays a key role in evaluation and improvement (Tuytens & Devos, 2010; Tuytens & Devos, 2014; Bryk & Schneider, 2002; Bryk et al., 2010). Teachers are more willing to recognize their weaknesses and try new instructional approaches when school leaders establish a culture of continuous improvement (Herlihy et al., 2014). Teachers are also more likely to respond to feedback in productive ways when they perceive the feedback as valid and worth acting on, otherwise they may ignore it or implement it in superficial ways (Lane, 2020; Spillane et al., 2002).

Delivering Feedback within Teacher Evaluation Systems

Classroom observations are a near universal feature of new teacher evaluation systems. In the 2015-16 school year, 88% of public school teachers reported being formally observed and receiving instructional feedback at least once (Tuma et al., 2018). Observation rubrics commonly used in teacher evaluation systems provide a common language for discussing high-quality instruction (Kraft & Gilmour, 2016). However, most new evaluation systems require very few observations and do not mandate post-observation meetings where evaluators discuss their feedback with teachers in-person (Steinberg & Donaldson, 2016). Often feedback is limited to a single formal written evaluation that teachers receive at the end of the year.

Research suggests that districts have implemented new evaluation systems in ways that undercut administrators' ability to provide frequent and effective feedback. Most states have added the responsibility of conducting time-intensive teacher observations to administrators' existing tasks without providing additional supports or training (Kraft & Gilmour, 2016; Neumerski et al., 2018). This can lead administrators to engage in satisficing behavior such as

conducting brief observations and providing generic positive feedback (Halverson et al., 2004). In Chicago, researchers found that administrators dominated post-observation conversations and rarely asked open-ended, higher-order questions that pushed teachers to reflect on their practices (Sartain et al., 2011). Principals who mainly see the evaluation process as an accountability tool report investing little time in providing feedback (Kraft & Gilmour, 2016). Thus, the quantity and quality of feedback teachers receive through the evaluation process is highly dependent on the skills, capacity, and goals of school leaders (Donaldson & Woulfin, 2018).

Several reports commissioned by federal and state education agencies have examined teachers' perspectives on evaluation feedback during the pilot phases of new evaluation systems (Donaldson et al., 2014; Cherasaro et al., 2016; Firestone et al., 2014). These studies found that teachers generally viewed their evaluation ratings as accurate and credible but had much more mixed responses about the timeliness and usefulness of the feedback they received. Most relevant to our work, Jiang, Sporte, and Luppescu (2015) analyzed teachers' perspectives on evaluation reforms in the first two years of implementation in Chicago Public Schools (CPS). Most teachers reported receiving feedback that was specific and actionable, but felt the process was stressful and not worth the time and paperwork it required.

The Effects of Feedback on Teacher Performance

Research suggests that regular feedback can be a key driver of high performance among teachers. Teachers who are observed regularly and receive frequent feedback are more likely to report improving their instructional practices (Tuma et al., 2018). Evidence from randomized field trials of teacher coaching programs as well as peer observation and feedback programs suggests these low-stakes forms of feedback can produce meaningful improvements in teacher instruction and student achievement (Kraft et al., 2018; Burgess et al., in press; Papay et al.,

2020). Frequent feedback is also a core practice of effective urban charter schools (Angrist et al., 2013; Dobbie & Fryer, 2013) and high-performing traditional public schools (Reinhorn et al., 2017).

However, we know much less about whether the feedback provided to teachers as part of high-stakes evaluation systems promotes teacher professional growth. Three studies that credibly identify the effect of evaluation reforms in large urban districts point to the potential of evaluation feedback to serve as an engine for teacher professional growth (Taylor & Tyler, 2012; Steinberg & Sartain, 2015; Dee & Wyckoff, 2015). However, it is unclear whether the student achievement gains found in these studies are a result of evaluation feedback itself or the incentives these systems placed on teachers to maximize their efforts.

Two recent studies of evaluation reforms point to the challenges of improving instructional practice via evaluation feedback at scale. Garet and his colleagues (2017) found that introducing frequent observation and feedback cycles as part of a low-stakes evaluation system across eight districts had some positive impacts on instruction and achievement. A comprehensive study of sweeping evaluation and human capital reforms implemented by three school districts and four charter management organizations found these reforms failed to drive changes in instruction or achievement (Stecher et al., 2018). The mixed effects of these reforms are likely the product of varying implementation quality across states and districts.

Teacher Evaluation in BPS

In 2011, the Massachusetts Board of Elementary and Secondary Education adopted a comprehensive educator evaluation system "designed first and foremost to promote leaders' and teachers' growth and development" (Boston Public Schools, 2012 p.5). The regulations detailed a five-step evaluation cycle in which educators self-assess their own practice, develop goals with

their principals, collect evidence of their progress towards these goals, are observed by principals, and participate in a summative evaluation process. Principals and members of the school administration serve as evaluators, conducting a minimum of one announced and four unannounced observations for pre-tenure teachers and one unannounced observation for tenured teachers. Evaluators are required to provide written feedback to teachers within five days of each observation and encouraged to have in-person post-observation conversations with teachers, but do not score individual observations using the rubric.

Unlike many evaluation systems that apply a weighted formula and pre-established score thresholds to determine teachers' overall summative rating (Steinberg & Kraft, 2017), the BPS system takes a holistic approach. Evaluators consider evidence from classroom observations, instructional artifacts, and progress towards teachers' self-identified professional practice and student learning goals. Test-based measures of teacher performance were proposed but never integrated into the formal evaluation system. Evaluators rate teachers' overall performance across the academic year using a rubric developed by the state and adapted by BPS. These ratings consist of an overall rating on a four-point scale ranging from *Unsatisfactory* to *Exemplary* as well as ratings on four specific domains: 1) Curriculum, Planning, and Assessment, 2) Teaching All Students, 3) Family and Community Engagement, and 4) Professional Culture. Administrators are instructed to use their professional judgement to choose the overall rating supported by the preponderance of evidence.

All teachers receive either a formative or a summative rating each year depending on the length of their evaluation cycle. Teachers rated as *Proficient* or *Exemplary* proceed on either a 1- or 2-year evaluation cycle of self-directed growth with one unannounced observation each year. Teachers who are rated as *Needs Improvement* or *Unsatisfactory* are placed on 120-day or

yearlong structured improvement plans requiring detailed prescriptions from evaluators along with one announced and two to four unannounced observations. Teachers who do not improve after being placed on a more structured plan are moved to a 30-, 60-, or 90-day improvement plan. Receiving a summative rating below *Proficient* while on an improvement plan triggers the dismissal process. Teachers access their ratings online and are required to sign off that they had received them.

Evaluator Training Intervention

We worked in partnership with BPS to develop the "Providing Effective Feedback" professional development training series for BPS evaluators. BPS recruited eight experienced district principals with reputations as strong instructional leaders to help tailor training materials developed by the New Teacher Center to the local context for pilot testing in the 2012-13 school year. The district then assigned school-based evaluator teams to attend the multi-day training in one of four semesters across the 2013-14 and 2014-15 school years. Principals decided which members of their administrative staff would serve as evaluators and attend the training as well as which teachers they would evaluate. Data suggest these assignments were relatively stable with 70% of evaluators returning to their school the following year and 46% of teachers having the same evaluator as the prior year, on average.

BPS was guided by a philosophy that teachers' ongoing learning is necessary to ensure student success and such learning is possible when evaluators are trained to provide frequent meaningful feedback, encourage reflection, and create opportunities for professional growth. The curriculum was grounded in learning theories about adult behavioral change with a focus on moving school leaders from a compliance orientation to a growth-based orientation (Kegan & Lahey, 2009; Knowles et al., 2012; Merriam, 2001). The training emphasized practical strategies

for prioritizing time for the observation and feedback process, creating a shared vision of effective practice, helping teachers to identify actionable steps for improvement, and communicating with teachers in ways that promote positive interpersonal relationships and mutual trust.

The training taught evaluators how to conduct classroom visits, differentiate feedback based on teachers' needs, use coaching language, and develop a plan for supporting teacher growth through targeted professional development. Several sessions were specifically focused on teaching evaluators to provide clear, specific, and actionable feedback. For example, instead of simply telling a teacher they did something wrong, evaluators were encouraged to identify specific instances from classroom observations and help teachers address them using coaching language (e.g., "a few students were texting in class, consider using more group discussions to increase student engagement" vs. "students were off-task in class"). School leaders viewed and discussed videotaped lessons, practiced giving feedback through role-play, and debriefed about their experiences implementing feedback techniques in their own schools between sessions. The training also guided evaluators on how to use specialized conversation protocols to structure their feedback conversations (see Appendix Figure A1). Finally, the training provided evaluators with logistical tools (e.g., evaluation calendars) and time management strategies so they could conduct all the necessary classroom observations and discussion meetings.

There are several features that differentiated this training from other professional development courses: 1) the training was taught by BPS school leaders, who were doing the work that they were teaching about, instead of central office staff or external consultants; 2) the course was grounded in guiding philosophies and theories of adult learning, but also included practical strategies; 3) participants completed homework between sessions to practice what they

had learned; 4) participants received individualized feedback on their assignments; and 5) the training was intensive and occurred in small groups, consisting of 3-5 sessions totaling 15 hours with a cohort of approximately 20-30 peers. Training sessions typically occurred after school and were spread throughout the semester.

Methods

Sample

BPS, the largest school district in Massachusetts, enrolls almost 60,000 students across 123 schools including traditional public schools, charter schools, and pilot schools.[1] Using administrative records, we show in Table 1 that the majority of students are students of color and are designated as "high needs" by the district, with a sizeable percentage of English Language Learners (31%) and students with disabilities (21%).[2] Our analytic sample consists of the 4,805 teachers that were employed by the district during the 2013-14 and 2014-15 school years. As we show in Table 2, almost three-quarters of these teachers were female and approximately one-quarter held a graduate degree. About one-third of teachers were African-American or Hispanic.

A total of 355 evaluators – principals, vice principals, and other school leaders – worked in the 123 schools in our sample across both years. We report demographic information for these evaluators in Table A1. Like teachers, most evaluators were female (70%). Notably, a larger percentage of evaluators were persons of color compared to teachers (52% vs 39%). The typical evaluator had been in their current administrative position just over three years.

Randomization Design

Resource limitations required BPS to stagger the training program over the course of two years. This allowed us to randomize school-based evaluator teams to attend training sessions across four semesters (fall 2013, spring 2014, fall 2014, or spring 2015). We grouped eligible

schools into six blocks based on school size (above or below the median of 390 students) and type (elementary, middle, and high) and then randomized within the six school size-type blocks. We chose these blocks based on prior research that suggests there are systematic differences in school climate and teacher working conditions by school size and type (Herlihy et al., 2014; Lee & Loeb, 2000). We also hypothesized that administrators at larger schools may face more capacity constraints because they are assigned to evaluate more teachers. School-based evaluator teams could choose when during a semester to attend the sequence of training sessions, which were offered at three different times. In Table 1, we show that observable school characteristics are balanced across all four randomization groups.

Treatment-Control Contrast

In our primary analyses, schools randomly assigned to the fall 2013 and spring 2014 semesters serve as the treatment group and those assigned during the 2014-15 school year serve as the control group. This treatment-control contrast identifies the effect of being randomly assigned to attend the evaluator training program relative to business as usual in the district. In 2012-13, all BPS administrators were required to attend a two-day training that focused on introducing them to the new evaluation system. The training familiarized administrators with the new evaluation rubric and on-line evaluation system as well as helped to calibrate administrators' scores. This technical training did not address any elements of the evaluation feedback process. Thus, the treatment-control contrast we identify is between administrators offered the opportunity to receive targeted training on evaluation feedback verses the type of general introductory training for conducting evaluations that was typical in most districts.

Survey Instruments

Independent Teacher Survey. We administered a confidential, but individually identifiable survey to teachers, independent from the district administered teacher survey, to capture their views on the evaluation process at the end of the second (2013-14) and third (2014-15) years of the new teacher evaluation system. The survey consisted of 29 items measured on a five-point Likert scale. We developed the survey in collaboration with the BPS Office of Human Capital to ensure questions were aligned with both research purposes and district priorities.[3] Questions probed teachers' perceptions about evaluators' communication, fairness, utility, feedback, and relationship quality. We added items to capture more concrete evaluation implementation information such as the number of unannounced/announced observations evaluators made, the number of post-observation meetings evaluators and teachers had, and the length of these meetings.

Response rates for our independent teacher survey were 56% in 2013-14 and 60% in 2014-15.[4] These rates compare favorably to the Institute for Education Sciences' National Teacher and Principal Survey, which achieved a response rate of 57% in 2015-16 (Taie & Goldring, 2017). A broad range of teachers completed the survey, although survey respondents differed from non-respondents in several ways as shown in Table 2. For example, in 2013-14 teachers who completed the survey were more likely to be female (77% vs. 70%), white (64% vs. 57%), older (by less than a year), more experienced, and hold a graduate degree (28% vs 20%). Survey response rates also differed to a modest degree by teachers' evaluation ratings.

We administered the survey in June to maximize the probability teachers had received feedback prior to taking it. Teachers were observed and received feedback throughout the school year but did not receive their summative evaluation ratings until May or June. Knowing their evaluation rating may have influenced teachers' willingness to respond to the survey. We find

that teachers who took the survey in 2013-14 were somewhat less likely to have received an *Unsatisfactory* rating (1% vs. 2%), and more likely to have received a rating of *Exemplary* (19% vs. 14%) than those who did not. These patterns persisted in 2014-15.

We find no evidence that teachers' response rates were related to the timing of when they were randomly assigned to attend the training series. Thus, differential response rates do not pose a threat to the internal validity of our randomized field experiment. A second possible concern might be that teachers' survey responses were influenced by knowing their summative evaluation ratings. In our Robustness Tests section below, we show that our correlational analyses are quite consistent across a range of modeling specifications and alternative weights suggesting these threats are not major concerns.

BPS Climate Survey. BPS administers an anonymous annual school climate survey to teachers which asks about their school leadership and work environment, classroom instruction, classroom management, autonomy, and engagement and relationships with parents and students. The survey consists of 62 questions measured on a four-point Likert scale. BPS uses the survey to create school climate reports to inform families' school choice preferences and support school improvement efforts. Response rates were 69% and 79% in 2013-14 and 2014-15, respectively.

Independent Evaluator Surveys. We administered surveys to evaluators both at the beginning and end of the training series. The survey consisted of fourteen questions on a five-point or nine-point Likert scale covering a range of topics including evaluators' opinions about the evaluation system, the quality of the training, and their own ability to provide constructive feedback (see Appendix Table A2 for full items). We also asked evaluators to estimate the amount of time they spent observing teachers and analyzing data, writing evaluations, discussing

feedback, and setting goals. Across both years, 94% of evaluators who attended at least one training session completed the baseline survey and 88% completed the end-of-training survey.

Primary Outcomes

We examine a range of proximal, intermediate, and distal outcomes that we hypothesized would be affected by the feedback training program through a causal cascade of effects. Each of these measures aligns with a conceptual element in our theory of action described above.

Perceived Feedback Quality. We use eight items from our independent teacher survey to create a latent measure of teachers' perceptions about the quality of evaluation feedback they received. Together, these eight items have an alpha reliability of 0.95. Examples include, "how effective was your evaluator at communicating his/her feedback?" and "how much has your instruction improved because of the feedback you received from your evaluator?" (See Appendix Table A3 for all items). A principal component analysis suggests these eight items capture one primary principal component which explains 74% of the variance across items. We take the first principal component from a factor analysis and standardize it to have a mean of 0 and standard deviation (SD) of 1. Supplementary analyses on the construct and predictive validity of this measure presented in Appendix B show that it is correlated with a range of measures capturing the frequency and intensity of feedback as well as teacher improvement. For example, it is positively correlated with the number of times a teacher is observed, the frequency of post-observation feedback meetings, the length of meetings, how quickly evaluators meet with teachers after observing them, and improvements in teacher instruction as measured by their evaluation ratings. We also use these frequency measures of the feedback process as outcomes.

Classroom Instruction. We use ratings on two specific summative evaluation domains – *Curriculum, Planning, and Assessment* and *Teaching All Students* – to measure the quality of

teachers' classroom instruction. These instructional domains capture teachers' capacity to design and provide high-quality instruction that keeps students engaged, meets students' diverse needs, and fosters safe and collaborative learning environments. We view these as more exploratory outcomes given that it is possible the training program changed administrators' subjective approach to rating teachers' performance rather than teachers' underlying instruction.

Teacher Self-efficacy. We use the BPS school climate survey to measure teachers' self-efficacy for instructional strategies and classroom management. We measure self-efficacy for instructional strategies using three items (α = 0.85) and self-efficacy for classroom management using six items (α = 0.83). For each domain, we predict the first component from a factor analysis and standardize it to have a mean of 0 and SD of 1 (See Appendix Table A4 for items).

Student Achievement. We measure student achievement using student test scores from the Massachusetts Comprehensive Assessment System (MCAS). The MCAS is a statewide exam administered to students in grades 3 through 8 and 10 in math and ELA. We standardize scores at the year, grade, and subject level to have a mean of 0 and SD of 1.[5]

Analyses

Experimental Analysis. We estimate the intent-to-treat (ITT) effects of being randomly assigned to attend the evaluator training program during the 2013-14 school year on teacher-level and student-level outcomes using the following generalized ordinary least squares model for teacher (or student) *i* at school *s*:

$$Y_{is} = \alpha + \beta Treat_s + \delta X_{is} + \pi_b + \varepsilon_{is} \qquad (1)$$

The outcome Y_{is} represents either a teacher-level outcome such as the perceived quality of evaluation feedback or a student-level outcome such as achievement. $Treat$ is an indicator for being randomly assigned to the training program in the first year it was offered. For each

outcome, we present results from baseline models as well as from models that include controls to increase precision and test the sensitivity of our findings.

For teacher-level outcomes, we control for teacher, evaluator, and school characteristics. Here, X includes teacher and evaluator characteristics such as age, experience, gender, race, and education level.[6] School characteristics include total enrollment, student-to-teacher ratio, and percentages of students by race, high-needs students, English language learners, and students with disabilities as well as measures of eight school climate survey domain scores from the prior year. For student-level outcomes we control for student race, gender, special education status, eligibility for free or reduced price-lunch, grade level, and prior achievement. Across all specifications, we include school size-type blocks, π, to account for the stratified randomization process and cluster standard errors at the school level.

We extend these ITT analyses by also estimating the treatment-on-the-treated (TOT) effects of attending the training using two-stage least squares. For the first set of analyses, we use a binary indicator to compare those that attended at least one session to those that attended no sessions. For the second set of analyses, we use a continuous measure of the proportion of training sessions attended to identify rigorous experimental estimates of the effect of attending all sessions. In both these analyses, the random assignment to attend the training serves as our instrument to isolate exogenous variation in these measures of program attendance. We also estimate two additional treatment-control contrasts using equation (1). We examine the effect of being randomly assigned to attend the training in the fall verses the spring to assess if receiving training earlier in the year had a larger effect on evaluators' practices at the end of the year. For these analyses, we pool results across both years and compare outcomes at the end of the first year for those assigned in fall 2013 to those assigned in spring 2014, and outcomes at the end of

the second year for those assigned in fall 2014 to those assigned in spring 2015. We also include year fixed effects to account for variation in outcomes across years.

Finally, we estimate the medium-term effect of the training one year later by defining the treatment group as evaluators randomized to attend sessions in fall 2013 and spring 2014 and the control group as evaluators randomized to attend during spring 2015 and compare outcomes in spring 2015. This allows us to test the lasting power of the training by examining if evaluators who were trained in the first year improve outcomes in the year after they completed the training relative to evaluators who had just completed the training. We omit schools assigned to treatment in the fall of 2014 from these analyses. This third treatment-control contrast serves as a lower-bound estimate of any medium-term effects given the control group received the offer for training as well.

Predictive Analysis. To answer our third research question about the predictors of perceived evaluation feedback quality, we construct a teacher-year level dataset that links individual teachers to their evaluators across the 2013-14 and 2014-15 school years. We model perceived evaluation feedback quality for teacher *i* at school *s* in year *t* as follows:

$$Evaluation\ Feedback_{ist} = \alpha + \beta X_{ist} + \gamma_t + \varepsilon_{ist} \quad (2)$$

We include fixed effects for year, γ, and cluster standard errors at the school level. Here our vector of predictors includes the teacher, evaluator, and school covariates described above. We also include indicators for school type (elementary, middle, or high), indicators for the overall summative rating category teachers received, indicators for teacher licensure endorsement areas (e.g., math, science), and a count of the number of teachers an evaluator observes in a given year.

Findings

RQ1: Teachers' Perceptions about the Evaluation System and Performance Feedback

We begin by describing teachers' general experiences with the BPS evaluation system to provide a better understanding of the context in which the district implemented the feedback training program. For these descriptive analyses, we draw on all BPS teachers who responded to our independent survey in the two years the district implemented the training program, 2013-14 and 2014-15. These represent the second and third years of district-wide implementation of the new evaluation system, respectively. The pattern of findings we report here are unchanged if we narrow our focus to only those teachers whose schools were randomly assigned to the control group in the first year of the training program and would not have been affected by the treatment.

The Evaluation Context. Teachers' responses to our independent survey indicate that the district provided a fruitful context in which to test the efficacy of the feedback training for promoting teacher development. The necessary conditions of trusting relationships, frequent observations, and perceptions of valid ratings appeared to be largely established at most schools. BPS evaluators were successful at evaluating teachers regularly and differentiating their performance to some degree. Teachers reported that evaluators made, on average, 3.63 unannounced and 1.91 announced visits during the school year, well in line with the recommended number of teacher observations. Approximately 6% of teachers received overall ratings of *Unsatisfactory* or *Needs Improvement*, while 76% of teachers were rated *Proficient* and 18% were rated *Exemplary*. This distribution of ratings reflects a strong skew towards higher ratings, but also greater differentiation than prior to the evaluation reforms when 99.2% of BPS teachers were rated as "Meets or Exceeds Standard" (NCTQ, 2010). It also represents greater differentiation than most new teacher evaluation systems (Kraft & Gilmour, 2017). The two instruction-specific evaluation domains share a similar ratings distribution.

Most teachers believed that evaluators were fair and accurate, and they felt that they had a strong relationship with their evaluator. Almost 70% of teachers agreed that their evaluator's assessment of their performance was fair. Roughly two-thirds of teachers agreed that evaluators based their feedback on direct evidence and provided accurate assessments of their teaching. Furthermore, three-quarters of teachers agreed their relationship with their evaluator was characterized by mutual respect and about 60% said they trusted their evaluator and felt their evaluator was committed to supporting them to improve their teaching practices.

However, teachers held far less favorable views about the quality of feedback administrators provided as part of the evaluation process. In Figure 2, we show the percent of teachers who responded positively to the eight items used to measure perceived feedback quality. Only half of teachers surveyed said that they were satisfied with the quantity of feedback they received and less than half felt that their feedback was useful or actionable. Ultimately, just over a quarter of teachers felt that their instruction improved because of this feedback.

***Systematic Challenges*.** Several implementation challenges likely contributed to teachers' perceived lack of consistent, high-quality feedback. First, evaluators struggled to find time to meet with teachers and provide feedback. On average, evaluators conducted two announced observations and four unannounced observations, but only met with teachers in-person twice a year to provide feedback. One-third of teachers reported *never* meeting with an evaluator to have a post-observation feedback discussion.

Large evaluator loads were likely a factor that limited the frequency and length of post-observation meetings. A typical evaluator assessed a dozen teachers in a year. Approximately 17% of evaluators in 2013-14 and 10% in 2014-15 evaluated 20 or more teachers in a given year. In Figure 3, we show the distribution of how long teachers estimated a typical post-observation

meeting lasted ranging from only a few minutes to an hour. Teachers estimated that they met with evaluators for an average of 20 minutes, implying that evaluators spent an average of only 40 minutes a year discussing feedback with each teacher.

The design of the online evaluation data portal, which tracked written feedback but did not require principals to submit information about in-person debrief meetings, likely incentivized evaluators to focus their time on submitting written feedback rather than having feedback conversations. Prior to completing the training program, we asked evaluators to estimate how they allocated their time across different parts of the evaluation process. Evaluators reported spending most of their time observing and analyzing data (34%) and writing evaluations (28%) compared to time spent setting goals (18%) and discussing feedback (20%) with teachers. The more limited attention given to in-person feedback and development likely contributed to teachers' negative perceptions of feedback.

RQ2: The Effects of the Evaluator Training Program

***Program Implementation*.** Evidence suggests that the BPS evaluator training was attended modestly-well by administrators and widely viewed as valuable in both school years. Sixty percent of evaluators randomly assigned to attend the program in the first year attended at least one of the 3 to 5 training sessions, and 71% in the second year. Among evaluators that did attend, most attended several sessions or completed the full training. Ultimately, 40% of evaluators in BPS attended all the assigned training days in 2013-14 and 52% in 2014-15 (see Appendix Table A5 for more details).

Evaluators overwhelmingly rated the training they received favorably and felt it would help them improve their evaluation feedback. Figure 4 illustrates how evaluators felt more capable at providing high-quality evaluation feedback after completing the training. Using

measures on a 9-point Likert scale ranging from *novice* (1) to *expert* (9), we find that after completing the training series in the first year, evaluators rated themselves higher, on average, at identifying improvement areas (0.54 point increase), providing individualized feedback (0.54 point increase), communicating feedback effectively (0.46 point increase), and suggesting actionable steps for improvement (0.60 point increase). These self-assessed improvements were similar in the second year. Moreover, evaluators rated themselves as generally very satisfied with the training (means of 7.92 in year 1 and 7.17 in year 2 on a 9-point scale ranging from *not at all satisfied* (1) to *extremely satisfied* (9)) and felt that the quality of training they received was high relative to other BPS professional development programs (means of 7.77 in year 1 and 7.40 in year 2 on a 9-point scale ranging from *substantially worse* (1) to *substantially better* (9)). Evaluators also reported that they were likely to incorporate techniques from the training to better support the professional development of their teachers during future evaluations.

Intent-to-Treat Estimates. We find no effects of being randomly assigned to attend the training program in the first year on a range of proximal, intermediate, and distal outcomes identified in our theory of action. As shown in Table 3, our estimate of the ITT effect on perceived evaluation feedback quality is small, negatively signed, and statistically insignificant (-0.02 SD). In our preferred model, we do find evidence that being assigned to attend the training had a marginally significant effect on the time between observations and post-observation feedback meetings, reducing it by 1.30 days relative to the control group mean of 5.34 days. However, we fail to find any significant effects on other implementation measures such as the number of observations or the number and length of post-observation meetings. We also find that the training had no effect on teachers' instructional effectiveness as measured by their evaluation ratings on instructional domains.

Given the lack of positive effects on more proximal outcomes, it is not surprising that we do not find positive effects on teacher self-efficacy or student achievement. We estimate effects on student achievement in two samples, our full analytic sample of over 54,000 students and a subsample of over 42,000 students that also have MCAS scores from the prior year. Including lagged test scores serves to substantially increase the precision of our estimates despite the reduced sample size which excludes 3rd graders. Focusing on our estimates with lagged prior achievement controls, we can rule out effects as small as 0.04 SD in math and 0.09 SD in ELA.

The intervention appears to have had, if anything, a moderate negative effect on teachers' self-efficacy for classroom management and instructional strategies. Our preferred results suggest that assigning administrators to attend the training caused a significant 0.20 SD reduction in teachers' self-efficacy for classroom management and a marginally significant 0.19 SD decrease in teachers' self-efficacy for instructional strategies.

Treatment-on-the-Treated Estimates. In Appendix Table A6, we provide TOT estimates of attending at least one training session and the proportion of training sessions attended. Conceptually, the estimates from attending at least one session rescale those in Table 3 by dividing by the take-up rate of 0.6. Our second set of estimates examines the effect of attending all sessions (i.e. where is one unit change is moving from 0% to 100%). The general pattern of substantively small and non-significant results remains. These results provide little evidence that the modest attendance rates of the training program are driving our null results.

Alternative Treatment-Control Contrasts. We find no differential ITT effects based on the timing an evaluator attended the training during the school year on almost any of our outcomes of interest (see Appendix Table A7). We also find null or negative ITT effects of the training program on outcomes measured the year after evaluators completed the training. This

suggests that our primary treatment effects are not likely to be biased downward because half of the evaluators only completed the training towards the end of the spring semester. We again find a negative and statistically significant impact on teachers' self-reported classroom management practices and instructional strategies. We also find a marginally significant 0.07 SD negative effect on student achievement in ELA. Given the number of outcomes we examine across our three treatment-control contrasts, it is also possible that these negative achievement impacts are spurious. Common approaches to adjusting for multiple-hypothesis testing reinforce our overall conclusion that the training had little to no effect on the range of outcomes we measure.

RQ3: Predictors of High-Quality Evaluation Feedback

Given teachers' mixed experiences with evaluation feedback and the limited success of the training program, we seek to inform future efforts to improve evaluation feedback through a range of exploratory analyses. All of the main results we report below draw on our full sample of BPS teachers who completed the evaluator survey in both years.

What does effective feedback look like? In Table 4, we disaggregate teachers' responses to a range of questions about their experience with evaluation feedback based on their perceptions about how much this feedback helped them to improve their instruction. Teachers who reported that feedback helped them improve a tremendous amount were observed by their evaluators almost twice as many times (7.74 vs. 3.98) and had more than twice as many post-observation meetings (4.73 vs. 2.08) compared to teachers who did not feel feedback helped at all. Their evaluators also met with them sooner after observing them (in 3.65 days vs. 6.01 days).

There were also stark differences in the perceived quality of feedback and relationships with evaluators across these two groups. Teachers who reported the largest improvements from feedback characterized their feedback conversations as opportunities for reflection, as providing

actionable feedback, and as part of a positive, trusting professional relationship. In contrast, almost every teacher that felt their evaluation feedback was unhelpful responded that the feedback they received was not actionable and that their evaluator did not ask them to reflect in-depth or assess their own teaching. The vast majority of these teachers also reported that their relationship with their evaluator was not built on trust.

Do evaluators differ in their perceived effectiveness? Our data reveal that some evaluators were far more effective at providing feedback that was perceived to be high-quality than others. As shown in Appendix Table A8, a variance decomposition of perceived feedback quality where teachers are nested within evaluators reveals that 16% of the variance occurs between evaluators. These estimates, however, conflate evaluator effects with any systematic differences across schools such as culture and climate that might affect the observation and feedback process. We next fit a multilevel model, nesting teachers within evaluators and evaluators within schools and find that schools only account for 4% of the variation, while differences in evaluators within the same school explain 13%. Thus, evaluators differ in their overall ability to provide feedback even within the same school.

In Figure 5, we plot the distribution of average evaluation feedback quality ratings across all evaluators who evaluated at least five teachers with survey data during the two years of the study. Across the 335 evaluators with sufficient data, 49 (15%) had average perceived feedback quality ratings that were statistically significantly above the mean, and 37 (11%) had ratings that were statistically significantly below the mean. This reflects far more variation in average feedback quality across evaluators than we would expect by chance, suggesting there are real differences in evaluators' ability to provide high-quality feedback.

Which evaluators are perceived to give higher-quality feedback? In Table 5, we explore which evaluator characteristics are predictive of high-quality feedback.[7] We find that the two most important evaluator characteristics are tenure at a school and race. Teachers rate evaluators with more experience at their school as providing higher-quality feedback. Compared to evaluators with 0-2 years of tenure at their school, evaluators with 6-8 years of tenure are reported to provide feedback that is 0.19 SD higher quality. Teachers also rate evaluators of color as providing lower-quality feedback. For example, a teacher that has an African-American evaluator compared to a white evaluator is likely to report that their evaluation feedback quality is almost a quarter of a SD lower. These patterns raise questions about whether some teachers are less receptive to feedback, or more critical of it, when it comes from evaluators of color. We find evidence of a weak negative association between the number of teachers an evaluator observes and perceived feedback quality that is only significant in some models.

A growing literature now documents the importance of racial congruence between teachers and students for student outcomes (Dee, 2004, 2007; Egalite et al., 2015; Lindsay & Hart, 2017; Holt & Gershenson, 2015; Gershenson et al., 2018). Given the sizable samples of teachers of color (39%) and evaluators of color (52%) in our study, our data present the opportunity to better understand how evaluator and teacher racial congruence is related to perceived feedback quality. We find that racial congruence in teacher and evaluator pairs is associated with positive perceptions of evaluation feedback quality among teachers of color. When African-American teachers have an African-American evaluator, they report receiving feedback that is about 0.30 SD higher than racially incongruent pairs.[8] We find positive estimates of similar magnitudes for racial congruence among Hispanics and Asians of 0.30 SD and 0.34 SD, respectively, although the estimate for Asians is not statistically significant.

Robustness Tests. We conduct several robustness checks to examine if the racial congruence and other patterns we report above remain consistent across alternative model specifications. It is possible that the relationships we observe reflect teachers' satisfaction with their summative evaluation rating since most teachers filled out surveys after receiving their summative ratings. Results from column 1 and 2 of Table 5 illustrate how our results are unchanged regardless if we exclude or include controls for teachers' summative rating. In column 3, we restrict our sample to the 76% of teachers who received the same rating of *Proficient* and find strikingly similar results.

In columns 4 and 5, we show that our estimates remain quite similar when we restrict our comparisons to teachers within the same school or even with the same evaluator by including school or evaluator fixed effects. This suggests that time-invariant school characteristics such as school size and grade level, and dynamic teacher-evaluator sorting patterns are not driving our findings. In column 6, we find similar results when we apply weights based on the differential school-level response rates to our teacher survey. Finally, we also find similar estimates in column 7 after inversely weighting by teachers' propensity to complete our survey suggesting that our results would have been similar had all teachers responded.

Discussion

Can teacher evaluation systems produce high-quality feedback?

The findings from our descriptive, experimental, and exploratory analyses provide evidence in support of a nuanced answer. We find large systematic differences across evaluators in the perceived quality of feedback they provide. At the same time, we find that most teachers do not report receiving evaluation feedback that helped them improve their instructional practice and that a training program designed specifically to improve evaluation feedback was largely

ineffectual. While it is possible for administrators to provide feedback that teachers view as high-quality, most administrators currently do not meet this goal. Furthermore, administrator training programs alone do not appear to be the solution for improving evaluation feedback at scale.

The evaluator training program we studied was designed by experienced district-based evaluators, was grounded in adult learning practices, prioritized active learning, provided specific tools and techniques to conduct effective discussion meetings, and was pilot tested. Moreover, evaluators liked the training, thought it was of high quality, and intended to use practices learned during the evaluation process. However, the disappointing impacts of the program are consistent with several prior randomized controlled trials of principal training programs that aimed to improve their instructional leadership and feedback skills (Goff et al., 2014; Herrmann et al., 2019; Jacob et al., 2015; Mihaly et al., 2018). The negative impacts we find on teachers' self-efficacy raise further concerns that administrators' critical feedback was too broadly focused on the teacher themselves rather than specific instructional practices (Kluger & DeNisi, 1996). It is also possible these findings reflect teachers recalibrating their own self-assessments to a higher standard rather than actual declines in instructional quality.

We see several possible explanations for this pattern of null results. Improving administrators' ability to provide feedback is only valuable if administrators can make time for feedback conversations. Prior research documents how capacity constraints limit evaluators' ability to find time to meet with teachers regularly (Kraft & Gilmour, 2016; Donaldson, 2012; Donaldson & Woulfin, 2018). During training sessions many BPS evaluators were candid about the fact that they would not have enough time to implement new feedback strategies with all the teachers they were responsible for evaluating. Evaluators were, on average, conducting 50-60 required classroom observations a year and many more brief observations. Time constraints may

have also caused evaluators to prioritize elements of the evaluation system that district administrators could closely monitor such as evaluation ratings and written feedback.

The evaluator training program was also not designed to address evaluators' limited content knowledge and grade-level experience. Evaluation systems that require evaluators to assess and provide feedback to teachers across grades and subjects can result in feedback focused on general pedagogy instead of content-specific pedagogical knowledge (Kraft & Gilmour, 2016). Consistent with this, our exploratory analyses in Table 5 reveal that teachers who taught subject-specific classes reported receiving lower-quality evaluation feedback than their peers who were subject-generalists even when comparing teachers at the same school with the same evaluator.

Another possible explanation is that teachers are hesitant to fully engage in open conversations about their strengths and weakness when feedback is provided within the context of a high-stakes evaluation system. High-stakes evaluation systems can cause teachers to become guarded about their instructional practice for fear that their weakness might be used against them in the evaluation process (Lane, 2020). Several studies have documented how new high-stakes evaluation system can create tension and strain relationships between administrators and teachers (Donaldson, 2016; Neumerski et al., 2018). As our theory of action illustrates, improving teachers' instructional practices through high-quality evaluation feedback requires that teachers be willing and able to act on this feedback in productive ways (Danielson, 2000).

What can districts do to improve the quality of evaluation feedback teachers receive?

In our view, promoting professional development at scale via evaluation systems likely requires substantial investments in dedicated instructional leadership positions for this purpose rather than relying on thinly stretched administrators to drive instructional improvement. For

example, Cincinnati and D.C. Public Schools were able to drive meaningful improvements in teachers' instructional practices and student achievement using evaluation systems that employed experienced teachers and instructional experts as full-time evaluators (Taylor & Tyler, 2012; Adnot et al., 2017). Our analyses also suggest there are likely administrators in every district that are successful at providing high-quality feedback. Districts might prioritize hiring principals who are instructional experts and deploying its most effective evaluators across multiple schools via full-time instructional leadership or coaching roles. Investing in full-time evaluators or coaches would allow districts to better match teachers with instructional experts who have experience teaching the same content and grade level.

Our findings also point to specific practices that evaluators might adopt. Similar to prior studies (Tuma et al, 2018), we find that teachers were more likely to report that their feedback helped them improve when they were observed and met with their evaluators more frequently; when they were invited to be active participants in diagnosing their performance; when the feedback they received was actionable and based on direct evidence; and when their relationship with their evaluator was characterized by trust and mutual respect. Accomplishing this requires individual skill on the part of evaluators as well as an organizational commitment to prioritizing feedback conversations and protecting evaluators' time to make frequent conversation possible.

Our findings also illustrate the importance of developing a diverse corps of evaluators. Studies find large positive benefits for students of color when they are taught by a teacher who shares their same race (Dee, 2004, 2007; Egalite et al., 2015; Lindsay & Hart, 2017; Holt & Gershenson, 2015; Gershenson et al., 2018). We find analogously that teachers of color appear to benefit when their evaluators share their same race. The positive association between teacher-

evaluator racial congruence and perceived feedback quality speaks to the benefits of recruiting, developing, and retaining diverse school leaders for both teachers and students.

Study Limitations

Our analyses face several important limitations. Our primary measure of feedback quality is based on teachers' subjective perceptions rather than an objective assessment. It is encouraging, however, that these subjective perceptions are correlated with a range of objective measures we collected about the evaluation process. The timing of our surveys which we administered after most teachers had received their evaluation scores also creates the possibility that teachers' views of their feedback were colored by their performance ratings. While our results are robust to a range of alternative specifications, we cannot definitively disentangle this potential influence on perceived feedback quality.

We evaluate the training program in its first full year of implementation, which may not reflect program effectiveness in later years. BPS has remained committed to refining and improving the evaluator training series. We do find some evidence that teachers' perceptions of evaluation feedback improved slightly over time as shown in Figure 2. Another limitation is that our unit of analysis focuses on change at the individual evaluator level rather than examining school and system-wide structures which likely shape individuals' experiences. For example, work by Marsh et al. (2017) illustrates the important role of organizational contexts in shaping evaluation feedback quality. Finally, our findings are likely best generalized to other large urban school districts with similar evaluation contexts.

Conclusion

The passage of ESSA has provided states and districts with broad flexibility in how they evaluate teachers. States and districts looking to promote teacher development through their

evaluation systems should carefully consider the alignment between their stated goals, system design, and resource investments. Feedback in the form of evaluation ratings and one formal written assessment at the end of school year is unlikely to drive instructional improvement. Promoting teacher growth through evaluation feedback likely requires evaluators who are instructional experts with the time and skills necessary to provide frequent actionable feedback to teachers and actively involve them in assessing their own practice. Principals also play an important role by cultivating school cultures where teachers trust their evaluators and share a collective commitment to continuous improvement. States and districts that fail to invest in creating the systems and conditions that facilitate high-quality evaluation feedback are unlikely to succeed at promoting teacher development through the evaluation process.

Endnotes

1) Pilot schools are semiautonomous district schools that have autonomy over budgeting, staffing, governance, curriculum/assessment, and the school calendar.

2) A student is considered high needs if he or she is designated as either low income, economically disadvantaged, ELL, former ELL, or a student with disabilities.

3) We adapted items from a range of existing teacher surveys, including the New Teacher Center's Teacher Working Conditions survey and the University of Chicago Consortium on School Research's 5 Essentials survey.

4) We took several steps to increase survey participation rates. First, we worked with the BPS central office to enlist the help of teacher leaders to inform their peers about the survey and encourage them to have their voices heard. We attended several district-wide teacher leader meetings where we presented our research design and described the survey. We also sent all teacher leaders a $10 Amazon gift card as a thank you several weeks in advance of administering the survey. Second, we administered the survey online via Qualtrics and tracked individual participation. This allowed us to send individualized invitations and follow-up reminders to teachers who had not completed the survey. Third, we used incentives including several drawings for Amazon gift cards between $100 and $300 dollars and school-wide breakfasts for all schools that had response rates of over 70%.

5) Eight schools that are alternative education programs or exclusively serve students with disabilities are excluded from these analyses because none of their students take the MCAS.

6) We measure teacher experience using teachers' experience step on the BPS salary schedule that approximates the number of years a teacher has worked in the district.

7) We find that school-level characteristics are, overall, only weakly associated with perceived evaluation feedback quality (see Appendix Table A9), so we focus our discussion on teacher and evaluator characteristics.

8) To be precise, we include in our model interactions that test whether perceived quality of feedback differs systematically for teachers who share the same race as their evaluator, relative to teachers that do not share the same race, above and beyond any average differences of being a teacher and having an evaluator of a given race.

References

Adnot, M., Dee, T., Katz, V., & Wyckoff, J. (2017). Teacher turnover, teacher quality, and student achievement in DCPS. *Educational Evaluation and Policy Analysis*, *39*(1), 54-76.

Almy, S. (2011). *Fair to everyone: Building the balanced teacher evaluations that educators and students deserve*. Washington, DC: Education Trust.

Angrist, J. D., Pathak, P. A., & Walters, C. R. (2013). Explaining charter school effectiveness. *American Economic Journal: Applied Economics*, *5*(4), 1-27.

Bell, C., Jones, N., Lewis, J., Qi, Y., Stickler, L., Liu, S., & McLeod, M. (2016). Understanding Consequential Assessment Systems for Teachers: Year 1 Report to the Los Angeles Unified School District.

Boston Public Schools. (2012, March). The Boston Public Schools implementation guide for the educator evaluation system.

Bryk, A., & Schneider, B. (2002). *Trust in schools: A core resource for improvement*. Russell Sage Foundation.

Bryk, A. S., Sebring, P., Allensworth, E., Luppescu, S., & Easton, J. (2010). Organizing schools for improvement: Lessons from Chicago: University Of Chicago Press.

Burgess, S., Rawall, S., & Taylor, E. S. (in press). Teacher peer observation and student test scores: Evidence from a field experiment in English secondary schools. *Journal of Labor Economics*. Retrieved from https://www.journals.uchicago.edu/doi/pdf/10.1086/712997.

Center on Great Teachers and Leaders. (2014). *National picture: A different view*. Retrieved from http://www.gtlcenter.org/sites/default/files/42states.pdf.

Cherasaro, T. L., Brodersen, R. M., Reale, M. L., & Yanoski, D. C. (2016). Teachers' Responses to Feedback from Evaluators: What Feedback Characteristics Matter? REL 2017-190. *Regional Educational Laboratory Central*.

Curtis, R., & Wiener, R. (2012). *Means to an end: A guide to developing teacher evaluation systems that support growth and development*. Washington, DC: Aspen Institute.

Danielson, C., & McGreal, T. L. (2000). *Teacher evaluation to enhance professional practice*. ASCD.

Dee, T. S. (2004). Teachers, race, and student achievement in a randomized experiment. *Review of Economics and Statistics*, *86*(1), 195-210.

Dee, T. S. (2007). Teachers and the gender gaps in student achievement. *Journal of Human*

resources, *42*(3), 528-554.

Dee, T. S., & Wyckoff, J. (2015). Incentives, selection, and teacher performance: Evidence from IMPACT. *Journal of Policy Analysis and Management*, *34*(2), 267-297.

DeNisi, A. S., & Sonesh, S. (2011). The appraisal and management of performance at work. In S. Zedeck (Ed.), *APA handbooks in psychology. APA handbook of industrial and organizational psychology, Vol. 2. Selecting and developing members for the organization* (p. 255–279).

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

Desimone, L. M., & Garet, M. S. (2015). Best practices in teacher's professional development in the United States.

Dobbie, W., & Fryer Jr, R. G. (2013). Getting beneath the veil of effective schools: Evidence from New York City. *American Economic Journal: Applied Economics*, *5*(4), 28-60.

Donaldson, M. L. (2012). Teachers' Perspectives on Evaluation Reform. *Center for American Progress*.

Donaldson, M. L., Cobb, C., LeChasseur, K., Gabriel, R., Gonzales, R., Woulfin, S., & Makuch, A. (2014). An evaluation of the pilot implementation of Connecticut's system for educator evaluation and development. *Storrs, CT: Center for Education Policy Analysis*.

Donaldson, M. L., & Papay, J. P. (2015). An idea whose time had come: Negotiating teacher evaluation reform in New Haven, Connecticut. *American Journal of Education*, *122*(1), 39-70.

Donaldson, M. L. (2016). Teacher Evaluation Reform: Focus, Feedback, and Fear. *Educational Leadership*, *73*(8), 72-76.

Donaldson, M. L., & Woulfin, S. (2018). From Tinkering to Going "Rogue": How Principals Use Agency When Enacting New Teacher Evaluation Systems. *Educational Evaluation and Policy Analysis*, *40*(4), 531-556.

Egalite, A. J., Kisida, B., & Winters, M. A. (2015). Representation in the classroom: The effect of own-race teachers on student achievement. *Economics of Education Review*, *45*, 44-52.

Feeney, E. J. (2007). Quality feedback: The essential ingredient for teacher success. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas*, *80*(4), 191-198.

Firestone, W. A., Nordin, T. L., Shcherbakov, A., Kirova, D., & Blitz, C. L. (2014). New Jersey's Pilot Teacher Evaluation Program: Year 2 Final.

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

Garet, M. S., Wayne, A. J., Brown, S., Rickles, J., Song, M., & Manzeske, D. (2017). The Impact of Providing Performance Feedback to Teachers and Principals. NCEE 2018 4001. *National Center for Education Evaluation and Regional Assistance*.

Garubo, R. C., & Rothstein, S. W. (1998). *Supportive supervision in schools*. Greenwood Publishing Group.

Gershenson, S., Hart, C., Hyman, J., Lindsay, C., & Papageorge, N. W. (2018). *The long-run impacts of same-race teachers* (No. w25254). National Bureau of Economic Research.

Glickman, C. D. (2002). *Leadership for learning: How to help teachers succeed*. ASCD.

Goff, P., Edward Guthrie, J., Goldring, E., & Bickman, L. (2014). Changing principals' leadership through feedback and coaching. *Journal of educational administration*, *52*(5), 682-704.

Grissom, J. A., & Youngs, P. (Eds.). (2016). *Improving teacher evaluation systems: Making the most of multiple measures*. Teachers College Press.

Halverson, R., Kelley, C., & Kimball, S. (2004). Implementing teacher evaluation systems: How principals make sense of complex artifacts to shape local instructional practice. *Educational administration, policy, and reform: Research and measurement*, 153-188.

Hanushek, E. A. (2009). Teacher deselection. *Creating a new teaching profession*, *168*, 172-173.

Herlihy, C., Karger, E., Pollard, C., Hill, H. C., Kraft, M. A., Williams, M., & Howard, S. (2014). State and local efforts to investigate the validity and reliability of scores from teacher evaluation systems. *Teachers College Record*, *116*(1), 1-28.

Herrmann, M., Clark, M., James-Burdumy, S., Tuttle, C., Kautz, T., Knechtel, V., Dotter, D., Wulsin, C.S., & Deke, J. (2019). The Effects of a Principal Professional Development Program Focused on Instructional Leadership.

Holt, S. B., & Gershenson, S. (2015). The impact of teacher demographic representation on student attendance and suspensions.

Jacob, R., Goddard, R., Kim, M., Miller, R., & Goddard, Y. (2015). Exploring the causal impact of the McREL Balanced Leadership Program on leadership, principal efficacy, instructional climate, educator turnover, and student achievement. *Educational Evaluation and Policy Analysis*, *37*(3), 314-332.

Jiang, J. Y., Sporte, S. E., & Luppescu, S. (2015). Teacher perspectives on evaluation reform: Chicago's REACH students. *Educational Researcher*, *44*(2), 105-116.

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

Kegan, R & Lahey, L.L. (2009). *Immunity to change: How to overcome it and unlock potential in yourself and your organization*. Harvard Business Press.

Kimball, S. M. (2002). Analysis of feedback, enabling conditions and fairness perceptions of teachers in three school districts with new standards-based evaluation systems. *Journal of Personnel Evaluation in Education*, *16*(4), 241-268.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological bulletin*, *119*(2), 254.

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

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

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

Kraft, M.A., Blazar, D., Hogan, D. (2018). The effect of teaching coaching on instruction and achievement: A meta-analysis of the causal evidence. *Review of Educational Research, 88*(4), 547-588.

Lane, J. L. (2020). Maintaining the frame: Using frame analysis to explain teacher evaluation policy implementation. *American Educational Research Journal*, *57*(1), 5-42.

Lee, V. E., & Loeb, S. (2000). School size in Chicago elementary schools: Effects on teachers' attitudes and students' achievement. *American Educational Research Journal*, *37*(1), 3-31.

Lindsay, C. A., & Hart, C. M. (2017). Teacher race and school discipline. *Education Next*, *17*(1).

McLaughlin, M. W., & Pfeifer, R. S. (1988). *Teacher Evaluation: Improvement, Accountability, and Effective Learning*. Teachers College Press, Teachers College, Columbia University, New York, NY 10027.

Merriam, S. B. (2001). Andragogy and self-directed learning: Pillars of adult learning theory. *New directions for adult and continuing education*, *2001*(89), 3.

Mihaly, K., Schwartz, H. L., Opper, I. M., Grimm, G., Rodriguez, L., & Mariano, L. T. (2018). Impact of a Checklist on Principal-Teacher Feedback Conferences Following Classroom Observations. REL 2018-285. *Regional Educational Laboratory Southwest*.

National Council on Teacher Quality. 2010 *Human Capital in Boston Public Schools: Rethinking How to Attract, Develop and Retain Effective Teachers*. ERIC Clearinghouse.

Neumerski, C. M., Grissom, J. A., Goldring, E., Rubin, M., Cannata, M., Schuermann, P., & Drake, T. A. (2018). Restructuring Instructional Leadership: How Multiple-Measure Teacher Evaluation Systems Are Redefining the Role of the School Principal. *The Elementary School Journal*, *119*(2), 270-297.

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

Papay, J. P., Taylor, E. S., Tyler, J. H., & Laski, M. E. (2020). Learning job skills from colleagues at work: Evidence from a field experiment using teacher performance data. *American Economic Journal: Economic Policy*, *12*(1), 359-88.

Reinhorn, S. K., Johnson, S. M., & Simon, N. S. (2017). Investing in development: Six high-performing, high-poverty schools implement the Massachusetts teacher evaluation policy. *Educational Evaluation and Policy Analysis*, *39*(3), 383-406.

Sartain, L., Stoelinga, S. R., & Brown, E. R. (2011). *Rethinking teacher evaluation: Lessons learned from observations, principal-teacher conferences, and district implementation*. Chicago, IL: Consortium on Chicago School Research.

Scheeler, M. C., Ruhl, K. L., & McAfee, M. K. (2004). Providing performance feedback to teachers: A review. Teacher Education and Special Education, 27(4), 396–407.

Smither, J. W., London, M., & Reilly, R. R. (2005). Does performance improve following multisource feedback? A theoretical model, meta-analysis, and review of empirical findings. *Personnel psychology*, *58*(1), 33-66.

Spillane, J. P., Reiser, B. J., & Reimer, T. (2002). Policy implementation and cognition:

Reframing and refocusing implementation research. *Review of educational research*, *72*(3), 387-431.

Sporte, S. E., Stevens, W. D., Healey, K., Jiang, J., & Hart, H. (2013). *Teacher Evaluation in Practice: Implementing Chicago's REACH Students*. University of Chicago Consortium on Chicago School Research. 1313 East 60th Street, Chicago, IL 60637.

Stecher, B. M., Garet, M. S., Hamilton, L. S., Steiner, E. D., Robyn, A., Poirier, J., ... & de los Reyes, I. B. (2018). Improving Teaching Effectiveness.

Steinberg, M. P., & Sartain, L. (2015). Does teacher evaluation improve school performance? Experimental evidence from Chicago's Excellence in Teaching project. *Education Finance and Policy*, *10*(4), 535-572.

Steinberg, M. P., & Donaldson, M. L. (2016). The new educational accountability: Understanding the landscape of teacher evaluation in the post-NCLB era. *Education Finance and Policy*, *11*(3), 340-359.

Steinberg, M. P., & Kraft, M. A. (2017). The sensitivity of teacher performance ratings to the design of teacher evaluation systems. *Educational Researcher*, *46*(7), 378-396.

Taie, S., & Goldring, R. (2017). Characteristics of Public Elementary and Secondary School Teachers in the United States: Results from the 2015-16 National Teacher and Principal Survey. First Look. NCES 2017-072. *National Center for Education Statistics*.

Taylor, E. S., & Tyler, J. H. (2012). The effect of evaluation on teacher performance. *American Economic Review*, *102*(7), 3628-51.

Thomas, E., Wingert, P., Conant, E., & Register, S. (2010). Why we can't get rid of failing teachers. *Newsweek*, *155*(11), 24–27.

Thurlings, M., Vermeulen, M., Bastiaens, T., & Stijnen, S. (2013). Understanding feedback: A learning theory perspective. *Educational Research Review*, *9*, 1-15.

Tuma, A. P., Hamilton, L. S., & Tsai, T. (2018). A Nationwide Look at Teacher Perceptions of Feedback and Evaluation Systems.

Tuytens, M., & Devos, G. (2010). The influence of school leadership on teachers' perception of teacher evaluation policy. *Educational Studies*, *36*(5), 521-536.

Tuytens, M., & Devos, G. (2014). The problematic implementation of teacher evaluation policy: School failure or governmental pitfall? *Educational Management Administration & Leadership*, *42*(4_suppl), 155-174.

Zee, M., & Koomen, H. M. (2016). Teacher self-efficacy and its effects on classroom processes, student academic adjustment, and teacher well-being: A synthesis of 40 years of research. *Review of Educational research*, *86*(4), 981-1015.

Tables

Table 1. *School Characteristics Across Randomization Groups*

	Full Sample	Fall Year 1	Spring Year 1	Fall Year 2	Spring Year 2	P-value
Average Enrollment	513.99	510.30	501.77	509.59	534.86	0.99
Student to Teacher Ratio	12.27	12.88	11.22	12.59	12.42	0.16
Student Characteristics (%)						
Female	46.88	48.18	45.32	47.57	46.47	0.55
Race/ethnicity						
African-American	35.92	37.30	36.01	34.45	35.87	0.96
Asian	5.99	5.55	7.54	6.52	4.31	0.60
Hispanic	40.93	42.58	36.71	41.44	43.07	0.59
Other	2.44	2.25	2.36	2.92	2.23	0.43
White	12.57	11.91	13.73	14.02	10.61	0.75
High Needs[a]	83.53	84.37	83.93	80.86	84.91	0.57
English Language Learners	30.85	31.12	26.95	31.75	33.71	0.55
Students with Disabilities	20.64	20.71	22.34	21.36	18.10	0.73
Joint F-test ($\chi 2$ =7.80)						0.73
n	123	31	31	32	29	

Notes: All data is from SY 2012-13, pre-treatment. Year 1 refers to schools randomized to trainings during SY 2013-14 and year 2 refers to schools randomized to trainings during SY 2014-15. P-value are calculated from an F-test regressing treatment assignment (being randomly assigned in year 1 vs year 2) on school characteristics.

[a]A student is considered high needs if he or she is designated as either low income, economically disadvantaged, or ELL, or former ELL, or a student with disabilities.

Table 2. *Teacher Demographic Characteristics*

	2013-14				2014-15			
	All Teachers	Took Survey	Did not Take Survey	P-value	All Teachers	Took Survey	Did not Take Survey	P-value
Treatment[a]	51.02	51.39	50.56	0.59	51.18	52.14	49.76	0.13
Age	42.42	42.89	41.82	0.00	42.06	42.39	41.59	0.02
Female (%)	73.56	76.76	69.53	0.00	73.61	76.58	69.24	0.00
Graduate Degree (%)	24.82	28.28	20.46	0.00	23.40	26.90	18.22	0.00
Experience[b] (%)								
0-2	10.76	9.24	12.67	0.00	9.33	8.36	10.75	0.01
3-5	15.87	14.58	17.49	0.01	17.35	16.80	18.16	0.26
6-8	15.40	15.59	15.16	0.70	14.22	13.89	14.70	0.47
9+	57.98	60.59	54.69	0.00	59.11	60.95	56.39	0.00
BPS Summative Evaluation Rating	3.08	3.11	3.04	0.00	3.13	3.15	3.10	0.01
Rated "Unsatisfactory" (%)	1.49	0.96	2.22	0.00	0.95	0.59	1.56	0.00
Rated "Needs Improvement" (%)	5.54	5.17	6.06	0.23	3.64	3.61	3.70	0.89
Rated "Proficient" (%)	76.35	75.38	77.70	0.09	76.58	76.06	77.47	0.32
Rated "Exemplary" (%)	16.62	18.50	14.03	0.00	18.82	19.74	17.27	0.06
Race (%)								
African-American	21.98	19.20	25.49	0.00	21.08	18.70	24.61	0.00
Asian	6.12	5.76	6.57	0.27	6.07	6.22	5.85	0.63
Hispanic	10.05	10.08	10.02	0.94	10.17	10.26	10.04	0.82
Other	0.12	0.04	0.21	0.11	1.06	1.01	1.14	0.70
White	61.24	64.37	57.29	0.00	61.18	63.33	58.00	0.00
n	4,267	2,380	1,887		4,150	2,476	1,674	

Notes: Teacher demographic characteristics are calculated for teachers that did and did not take the independent teacher survey for SY 2013-14 and SY 2014-15. P-value are calculated via t-tests comparing demographic characteristics for teachers that took the survey and teachers that did not take the survey.

[a]Teachers from schools randomly assigned to training sessions in fall 2013 or spring 2014 (year 1) are in the treatment group and teachers from schools randomly assigned to training sessions in fall 2014 or spring 2015 (year 2) are in the control group.

[b]This variable takes discrete values corresponding to a teacher's years of experience teaching in the district (e.g., 7 corresponds to 7 years of teaching experience).

Table 3. *The Intent-to-Treat Effect of Evaluator Training on Teacher and Student Outcomes, Year 1 vs Year 2*

Outcomes	n	Uncontrolled	Controlled
Feedback Quality (PCA, standardized)	2,033	-0.07	-0.02
		(0.08)	(0.06)
Ever met to discuss feedback (Binary)	2,151	0.02	0.03
		(0.04)	(0.03)
Number of observations (Discrete)	2,094	0.17	0.32
		(0.38)	(0.38)
Number of discussion meetings (Discrete)	2,151	0.11	0.14
		(0.31)	(0.33)
Meeting length (Minutes)	2,151	0.19	-0.19
		(1.13)	(1.03)
Time between observation and meeting (Days)[a]	1,265	-0.61	-1.30*
		(0.66)	(0.76)
Summative Rating: Curriculum, Planning, and Assessment (Discrete, 1-4)	3,904	0.02	0.03
		(0.04)	(0.03)
Summative Rating: Teaching All Students (Discrete, 1-4)	3,904	0.02	0.03
		(0.04)	(0.03)
Self-efficacy for classroom management (PCA, standardized)[b]	2,907	-0.34***	-0.20**
		(0.12)	(0.09)
Self-efficacy for instructional strategies (PCA, standardized)[b]	2,907	-0.20	-0.19*
		(0.14)	(0.12)
Student math achievement[c]	53,664	-0.02	-0.00
		(0.14)	(0.08)
Student math achievement (Controlling for prior achievement)[c, d]	41,864	-0.04	-0.02
		(0.15)	(0.03)
Student ELA achievement[c]	53,056	0.03	0.05
		(0.12)	(0.06)
Student ELA achievement (Controlling for prior achievement)[c, d]	41,355	0.05	0.04
		(0.12)	(0.03)

Notes: *** p<0.01, ** p<0.05, * p<0.1. Standard errors in parenthesis are clustered at the school level. Controlled models use school characteristics—total enrollment, student to teacher ratio, percent of high needs students, percent of ELL students, percent of students with disabilities, and eight school climate survey domain scores from the prior year—and teacher and evaluator characteristics—age, experience, gender, race, and education.

[a]The sample is subset to teachers that ever met with an evaluator for a post-observation meeting.

[b]These outcomes are created by using the BPS school climate survey, which teachers answered anonymously. Since we cannot link individual teachers to their responses, we only control for school characteristics for these outcomes.

[c]Student achievement is measured in grades 3-8 and 10 in mathematics and ELA. We standardize scores at the year, grade, subject level to have a mean of 0 and standard deviation of 1.

[d]Since we include a lagged test score as a control, we exclude from the sample those who did not take the MCAS in the previous year (mostly third graders); this results in a loss of 22% of our sample.

Table 4. *How Teachers Described Observations and Feedback by How Much They Felt Feedback Helped Improve Instruction*

		How much teachers felt feedback helped improve instruction				
How teachers described observations and feedback	Full sample	Not at all	A little bit	Some	Quite a bit	A tremendous amount
Perceived feedback quality (Standard deviations)	0.00	-1.31	-0.52	0.35	0.99	1.47
Number of observations	5.52	3.98	4.74	5.87	6.84	7.74
Number of discussion meetings	3.38	2.08	2.89	3.24	4.04	4.73
Meeting length (Minutes)	19.70	19.66	17.67	19.67	20.44	21.88
Time between observation and meeting (Days)[a]	4.68	6.01	4.93	4.66	4.27	3.65
Percent of teachers agreeing[b]						
Asked to assess own teaching	29.55	3.67	11.13	32.51	58.84	78.19
Pushed to reflect in-depth	34.03	3.46	10.58	35.42	73.32	88.30
Received effective communication	53.27	9.17	29.21	65.09	91.87	95.70
Feedback based on direct evidence	65.39	25.20	46.40	79.69	94.20	97.87
Received accurate assessment	63.11	23.58	45.24	75.47	92.24	97.33
Received fair assessment	67.93	28.60	52.87	80.16	94.71	98.40
Received actionable recommendations	48.80	6.93	21.04	57.86	90.91	97.86
Received useful feedback to improve teaching	43.97	0.72	10.50	52.48	92.78	98.40
Satisfied with feedback quantity	49.85	6.66	20.97	61.70	91.24	96.28
Felt evaluator is committed to supporting improvement	59.40	12.63	36.36	75.32	94.68	97.86
Relationship with evaluator has mutual respect	75.40	39.69	63.03	88.87	96.52	99.47
Trust evaluator	61.65	20.55	41.36	75.96	92.24	96.81
Enjoy working with evaluator	62.74	21.75	42.46	76.82	93.37	98.93
n	8,417	983	948	1,504	986	188

Notes: Data are pooled from the SY 2013-14 and SY 2014-15 independent teacher surveys.

[a]The sample is subset to teachers that ever met with an evaluator for a post-observation meeting.

[b]Questions are on a five-point Likert scale. A teacher is considered agreeing with the statement if they answer in the top two choices (e.g., "agree" or "strongly agree").

Table 5. *The Relationship Between Teacher and Evaluator Characteristics and Perceived Evaluation Feedback Quality*

	Preferred model with no summative rating	Preferred Model	Teachers rated *Proficient* only	School fixed effects	Evaluator fixed effects	Weighted by teacher survey response rate	Weighted by propensity to take survey
Teacher characteristics							
Age	0.006***	0.010***	0.009***	0.011***	0.011***	0.010***	0.010***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Female	-0.113***	-0.151***	-0.141***	-0.153***	-0.128**	-0.152***	-0.149***
	(0.042)	(0.042)	(0.046)	(0.045)	(0.049)	(0.043)	(0.043)
Graduate degree	-0.078**	-0.086**	-0.081*	-0.071*	-0.076*	-0.086**	-0.078**
	(0.038)	(0.038)	(0.042)	(0.037)	(0.039)	(0.040)	(0.038)
Experience							
3-5	-0.047	-0.120	-0.129*	-0.110	-0.093	-0.114	-0.094
	(0.074)	(0.073)	(0.076)	(0.072)	(0.076)	(0.070)	(0.073)
6-8	-0.146*	-0.256***	-0.230***	-0.247***	-0.246***	-0.277***	-0.230***
	(0.078)	(0.078)	(0.084)	(0.075)	(0.076)	(0.079)	(0.078)
9+	-0.166**	-0.309***	-0.286***	-0.314***	-0.273***	-0.335***	-0.281***
	(0.075)	(0.076)	(0.086)	(0.074)	(0.079)	(0.076)	(0.075)
Race/ethnicity							
African-American	0.108	0.164**	0.220**	0.139*	0.108	0.176**	0.188**
	(0.077)	(0.076)	(0.092)	(0.073)	(0.079)	(0.075)	(0.080)
Asian	0.211**	0.246***	0.298***	0.228**	0.189**	0.246***	0.257***
	(0.094)	(0.087)	(0.096)	(0.088)	(0.089)	(0.091)	(0.087)
Hispanic	0.090	0.111	0.190**	0.116	0.061	0.121	0.123
	(0.080)	(0.077)	(0.085)	(0.075)	(0.084)	(0.077)	(0.079)
Endorsement[a]							
Math	-0.178***	-0.173***	-0.186***	-0.148**	-0.127*	-0.171***	-0.169***
	(0.061)	(0.060)	(0.068)	(0.062)	(0.072)	(0.060)	(0.059)
Reading	-0.172***	-0.180***	-0.219***	-0.193***	-0.163***	-0.183***	-0.178***
	(0.062)	(0.061)	(0.076)	(0.058)	(0.058)	(0.063)	(0.061)
Science	-0.129*	-0.138**	-0.136*	-0.128*	-0.083	-0.126*	-0.156**
	(0.073)	(0.069)	(0.072)	(0.073)	(0.070)	(0.072)	(0.068)
Social studies	-0.113**	-0.111*	-0.011	-0.091	0.007	-0.080	-0.101*
	(0.057)	(0.061)	(0.071)	(0.066)	(0.062)	(0.063)	(0.060)

Elementary	0.027	0.023	0.007	0.036	0.007	0.027	0.020
	(0.041)	(0.040)	(0.046)	(0.039)	(0.041)	(0.038)	(0.040)
Early childhood education	0.110*	0.093*	0.061	0.094*	0.053	0.108**	0.079
	(0.056)	(0.054)	(0.068)	(0.056)	(0.057)	(0.054)	(0.055)
Special education	-0.024	-0.025	-0.025	-0.032	-0.032	-0.033	-0.028
	(0.040)	(0.039)	(0.045)	(0.040)	(0.041)	(0.038)	(0.040)
English language learner	-0.076**	-0.090**	-0.079*	-0.083**	-0.067*	-0.089**	-0.095***
	(0.037)	(0.036)	(0.043)	(0.034)	(0.037)	(0.038)	(0.035)
Other	0.029	0.010	0.005	0.008	0.018	0.003	0.014
	(0.033)	(0.033)	(0.039)	(0.034)	(0.034)	(0.035)	(0.034)
Summative rating							
Needs improvement		0.377**		0.278	0.289	0.364**	0.359**
		(0.161)		(0.170)	(0.195)	(0.177)	(0.166)
Proficient		1.381***		1.246***	1.249***	1.393***	1.342***
		(0.166)		(0.173)	(0.199)	(0.179)	(0.173)
Exemplary		1.635***		1.511***	1.532***	1.646***	1.592***
		(0.173)		(0.179)	(0.206)	(0.186)	(0.177)
Evaluator characteristics							
Age	-0.007**	-0.009***	-0.010***	-0.008**		-0.008**	-0.009***
	(0.003)	(0.003)	(0.003)	(0.004)		(0.003)	(0.003)
Female	0.162***	0.148***	0.136**	0.145***		0.152***	0.145***
	(0.047)	(0.046)	(0.058)	(0.055)		(0.049)	(0.045)
Tenure at school							
3-5	0.036	0.005	-0.021	0.099*	0.051	-0.018	0.021
	(0.057)	(0.054)	(0.061)	(0.058)	(0.071)	(0.057)	(0.051)
6-8	0.197***	0.193***	0.153**	0.275***	0.033	0.184***	0.196***
	(0.062)	(0.059)	(0.074)	(0.068)	(0.120)	(0.061)	(0.058)
9+	0.174*	0.169	0.197	0.247**	0.025	0.130	0.197*
	(0.101)	(0.105)	(0.124)	(0.099)	(0.198)	(0.114)	(0.109)
Number of teachers evaluating	-0.003	-0.004	-0.006*	-0.007*	-0.009**	-0.004	-0.004
	(0.003)	(0.003)	(0.003)	(0.004)	(0.004)	(0.003)	(0.003)
Race/ethnicity							
African-American	-0.248***	-0.229***	-0.197**	-0.199*		-0.242***	-0.209**
	(0.081)	(0.081)	(0.092)	(0.101)		(0.088)	(0.083)
Asian	-0.312*	-0.311*	-0.436**	-0.450**		-0.349*	-0.296*
	(0.177)	(0.166)	(0.178)	(0.182)		(0.190)	(0.160)

Hispanic	-0.100	-0.100	-0.057	-0.173**		-0.082	-0.095
	(0.075)	(0.074)	(0.094)	(0.085)		(0.079)	(0.076)
Evaluator and teacher congruence							
Both same gender	-0.032	-0.006	-0.037	0.000	-0.031	0.000	-0.010
	(0.040)	(0.040)	(0.046)	(0.042)	(0.043)	(0.042)	(0.040)
Both African-American	0.305***	0.295***	0.276**	0.229**	0.273**	0.278***	0.275***
	(0.109)	(0.101)	(0.120)	(0.101)	(0.106)	(0.101)	(0.105)
Both Asian	0.441*	0.345	0.606*	0.255	0.255	0.344	0.355
	(0.248)	(0.237)	(0.334)	(0.236)	(0.270)	(0.254)	(0.233)
Both Hispanic	0.290***	0.294***	0.276**	0.356***	0.335**	0.277***	0.298***
	(0.109)	(0.103)	(0.122)	(0.110)	(0.131)	(0.100)	(0.105)
Both white	0.117	0.110	0.188**	0.081	0.062	0.110	0.133
	(0.079)	(0.078)	(0.093)	(0.077)	(0.085)	(0.082)	(0.081)
Survey response weights	N	N	N	N	N	Y	N
School fixed effects	N	N	N	Y	N	N	N
Evaluator fixed effects	N	N	N	N	Y	N	N
n	4,103	4,103	3,092	4,103	4,103	4,103	4,103

Notes: *** p<0.01, ** p<0.05, * p<0.1. Standard errors are in parenthesis.

Models use pooled data from SY 2013-14 and SY 2014-15 and estimate the relationship between teachers' perceived evaluation feedback quality and teacher, evaluator, and school characteristics (estimates for school characteristics are not shown in this table – for those estimates see Table A9). All models contain fixed effects for school year. Standard errors are clustered at the school level.

Dummy variables for race/ethnicity categories American-Indian and Native Hawaiian and Pacific Islander are also included but not reported in the table. The reference category is white.

[a] Endorsements are not mutually exclusive because teachers can be endorsed in multiple areas.

Figures

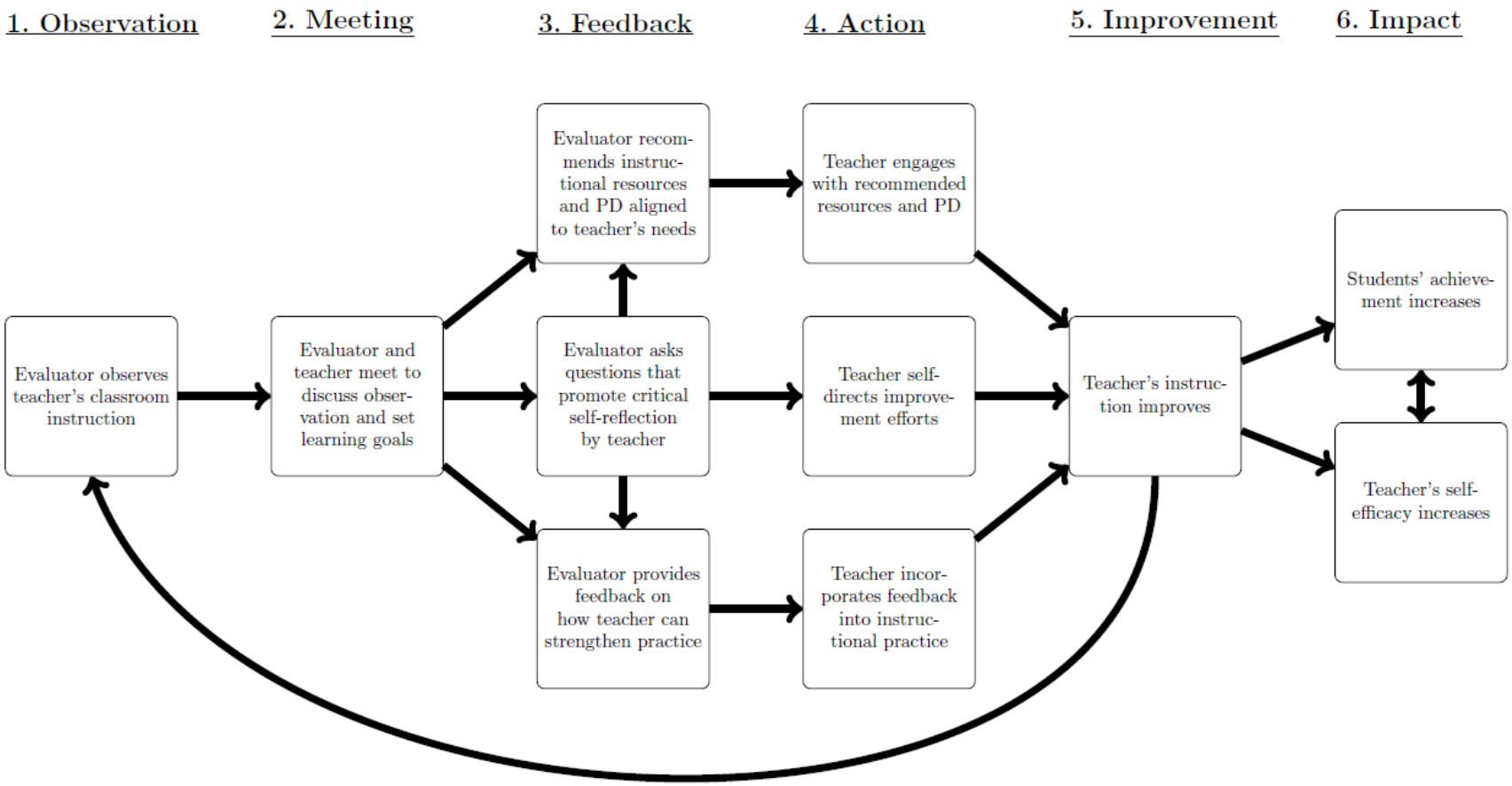

Figure 1. Theory of action behind observation and feedback cycles.

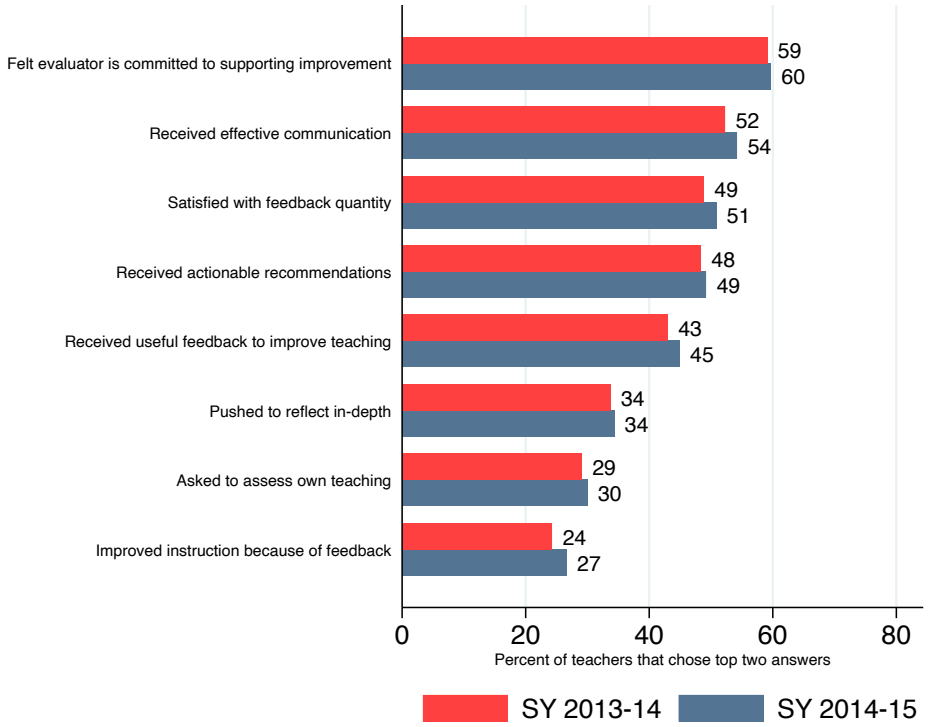

Figure 2. Agreement rates for items included in the perceived quality of evaluation feedback scale for the 2013-14 and 2014-15 school years.

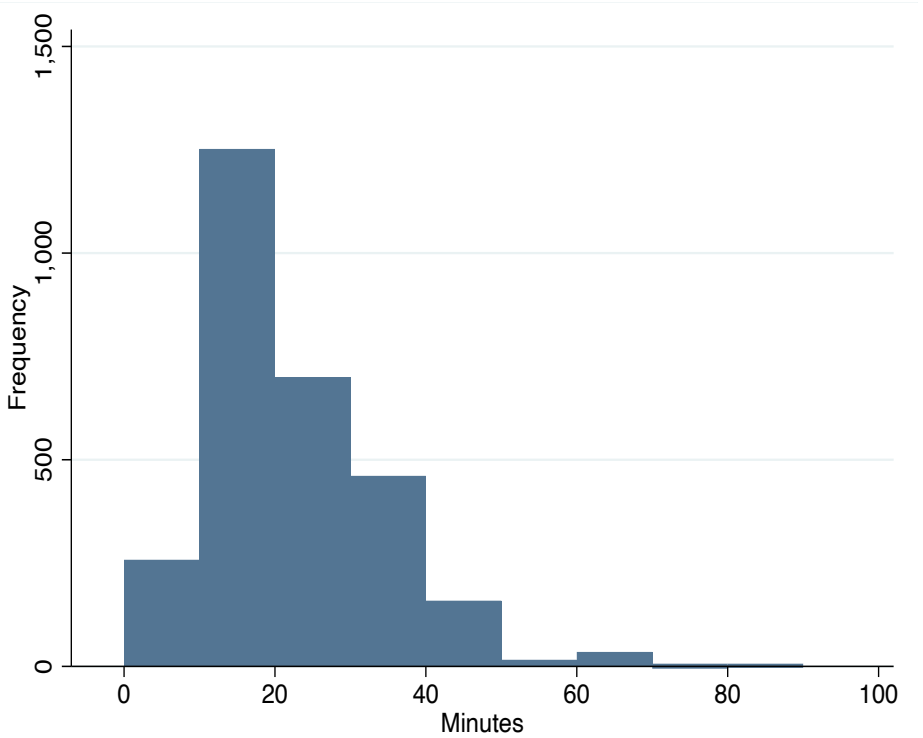

Figure 3. The length of post-observation meetings across the 2013-14 and 2014-15 school years.

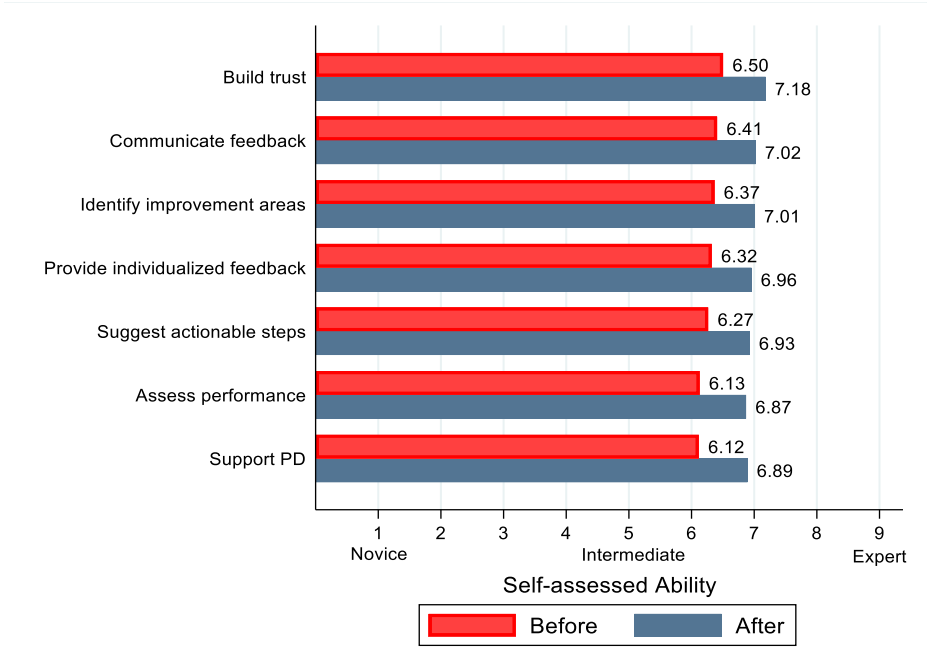

Figure 4. Evaluators' self-assessment of evaluation skills pre- and-post training pooled across the 2013-14 and 2014-15 school years.

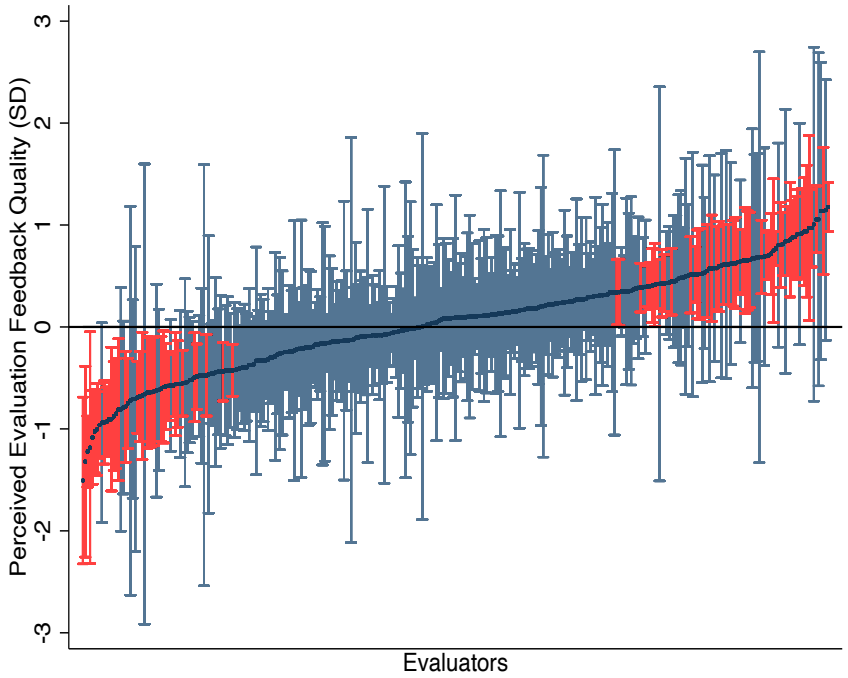

Figure 5. Distribution of average perceived evaluation feedback quality for evaluators pooling across the 2013-14 and 2014-15 school years.

Notes: This figure presents the average perceived feedback quality of evaluators in standard deviation units (SD), is subset to evaluators who evaluated at least five teachers, and only shows evaluators whose 95% confidence intervals are between -3 SD and 3 SD. This excludes 23 evaluators. Red confidence intervals denote evaluators whose perceived feedback quality was statistically significantly different from zero.

Appendix A

Table A1. *Evaluator Demographic Characteristics*

	2013-14				2014-15			
	All Evaluators	Did not attend any session	Attended any session	P-value	All Evaluators	Did not attend any session	Attended any session	P-value
Age	45.95	44.25	47.15	0.05	47.18	44.55	48.29	0.03
Female (%)	70.99	74.63	68.42	0.39	69.23	68.00	69.75	0.82
Number of teachers evaluating	13.12	10.36	15.10	0.00	10.54	9.47	11.00	0.18
Tenure at school (%)								
0-2	50.64	57.81	45.65	0.14	48.75	67.35	40.54	0.00
3-5	32.05	31.25	32.61	0.86	29.38	20.41	33.33	0.10
6-8	9.62	3.13	14.13	0.02	13.75	6.12	17.12	0.06
9+	7.69	7.81	7.61	0.96	8.13	6.12	9.01	0.54
Race (%)								
African-American	35.58	39.71	32.63	0.36	37.28	44.00	34.45	0.24
Asian	3.07	1.47	4.21	0.32	5.33	6.00	5.04	0.80
Hispanic	8.59	10.29	7.37	0.51	12.43	8.00	14.29	0.26
Other	0.02	0.04	0.00	0.04	0.01	0.02	0.00	0.12
White	50.92	44.12	55.79	0.14	44.38	40.00	46.22	0.46
n	177	70	107		178	51	127	

Notes: We calculate demographic characteristics for evaluators from SY 2013-14 and SY 2014-15 by those that attended no training session and any training session, regardless of whether or not the evaluator attended their assigned session. P-value calculated via t-tests comparing evaluators that attended any session to those that did not attend any session.

Table A2. *Items Included in Evaluator Pre- and Post-Training Survey*

Self-Assessment

1. Ability to assess teacher performance based on classroom observations using the BPS Educator Rubric Standards and Indicators.
2. Ability to identify the instructional practices that individual teachers most need to improve.
3. Ability to provide individualized feedback tailored to teachers' specific needs.
4. Ability to communicate feedback clearly and effectively.
5. Ability to suggest specific, actionable steps for teachers to meet their student-learning and professional-practice goals.
6. Ability to support the professional development of teachers through evaluation feedback.
7. Ability to use feedback conversations as opportunities to build trust and support.

Perspectives on Educator Evaluation

8. A primary purpose of the BPS Educator Evaluation System is to help teachers improve.
9. A primary purpose of the BPS Educator Evaluation System is to remove ineffective teachers from the classroom.
10. I am able to use the BPS Educator Evaluation System to help teachers improve.
11. I am able to use the BPS Educator Evaluation System to remove ineffective teachers from the classroom.

Overall Assessment of the Evaluator Training Program[a]

12. Overall, how likely are you are to incorporate techniques you learned in this evaluator training program in your own evaluation practices on a scale?
13. Overall, how satisfied you are with the evaluator training program?
14. Compared to other BPS professional development activities that you have participated in over the past three years, how would you rate the quality of the evaluator training program?

Notes: Questions are answered on a five-point or nine-point Likert scale by evaluators both before and after their training in SY 2013-14 and SY 2014-15.

[a]Included in post-training survey only.

Table A3. *Perceived Evaluation Feedback Quality Questions*

1. How often did your evaluator ask you to assess your own teaching during the evaluation?
2. How often did your evaluator ask you questions that pushed you to reflect in-depth?
3. How effective was your evaluator at communicating his/her feedback?
4. How actionable were your evaluator's recommendations about what you could do to improve your teaching?
5. How useful was your evaluators' feedback in supporting you to improve your teaching?
6. To what extent are you satisfied with the quantity of feedback you receive from your evaluator?
7. How much has your instruction improved because of the feedback you received from your evaluator?
8. How committed is your evaluator to supporting you to improve your teaching?

Notes: Questions are answered on a five-point Likert scale by teachers after their evaluation in SY 2013-14 and SY 2014-15.

Table A4. *Items Included in Teacher Self-efficacy Scales from BPS School Climate Survey*

Self-efficacy for instructional strategies
1. I can provide an alternative explanation or example when students are confused.
2. I can use a variety of assessment strategies in my class.
3. I can craft good questions for my students.
Self-efficacy for classroom management
1. How much can you do to control disruptive behavior in the classroom?
2. How much can you do to motivate students who show little interest in schoolwork?
3. How much can you do to get students to believe they can do well in schoolwork?
4. How much can you do to help your students value learning?
5. How much can you do to assist families in helping their children do well in school?
6. How much can you do to provide appropriate challenges for students who are excelling?

Notes: Questions from the BPS school climate survey are answered on a five-point Likert scale by teachers in SY 2013-14 and SY 2014-15.

Table A5. *Evaluator Training Attendance*

	Fall Year 1	Spring Year 1	Fall Year 2	Spring Year 2	Year 1[a]	Year 2[a]
Attendance to ANY assigned meeting in period	0.52	0.57	0.58	0.60	0.60	0.71
Attendance to 60% or more of assigned meetings in period	0.51	0.57	0.56	0.59	0.60	0.69
Attendance to ALL assigned meetings in period	0.31	0.41	0.46	0.43	0.40	0.52
Attendance percentage (of assigned meetings in period)	0.46	0.54	0.55	0.55	0.55	0.65
n	81	96	95	83	177	178

Notes: Attendance data is from SY 2013-14 and SY 2014-15.

[a]If an evaluator was unable to attend training sessions in a particular semester or missed multiple sessions, they were encouraged to attend sessions in a different semester, typically within the same school year. Therefore, attendance rates aggregated to the year level are slightly higher than at the semester level.

Table A6. *The Effect of Evaluator Training on Teacher and Student Outcomes (TOT), Year 1 vs Year 2*

		Attend At Least One Training Session		Proportion of Training Sessions Attended	
Outcomes	n	Uncontrolled	Controlled	Uncontrolled	Controlled
Feedback Quality (PCA, standardized)	2,033	-0.11	-0.03	-0.12	-0.03
		(0.12)	(0.10)	(0.13)	(0.10)
Ever met to discuss feedback (Binary)	2,151	0.03	0.04	0.04	0.05
		(0.06)	(0.05)	(0.06)	(0.06)
Number of observations (Discrete)	2,094	0.26	0.51	0.28	0.54
		(0.58)	(0.60)	(0.63)	(0.64)
Number of discussion meetings (Discrete)	2,151	0.17	0.23	0.19	0.25
		(0.46)	(0.53)	(0.51)	(0.56)
Meeting length (Minutes)	2,151	0.29	-0.31	0.32	-0.33
		(1.70)	(1.61)	(1.86)	(1.73)
Time between observation and meeting (Days)[a]	1,265	-0.91	-2.06*	-1.00	-2.22*
		(0.98)	(1.17)	(1.07)	(1.26)
Summative Rating: Curriculum, Planning, and Assessment (Discrete, 1-4)	3,904	0.03	0.05	0.03	0.05
		(0.06)	(0.04)	(0.06)	(0.04)
Summative Rating: Teaching All Students (Discrete, 1-4)	3,904	0.04	0.04	0.04	0.04
		(0.06)	(0.05)	(0.07)	(0.05)
Self-efficacy for classroom management (PCA, standardized)[b]	2,907	-0.41***	-0.25**	-0.54**	-0.33**
		(0.15)	(0.11)	(0.21)	(0.14)
Self-efficacy for instructional strategies (PCA, standardized)[b]	2,907	-0.25	-0.25*	-0.22	-0.25
		(0.17)	(0.15)	(0.25)	(0.20)
Student math achievement[c]	48,088	-0.07	-0.02	-0.07	-0.02
		(0.19)	(0.10)	(0.21)	(0.11)
Student math achievement (Controlling for prior achievement)[c, d]	37,461	-0.05	-0.04	-0.06	-0.05
		(0.04)	(0.03)	(0.04)	(0.04)
Student ELA achievement[c]	47,515	0.00	0.05	0.01	0.05
		(0.16)	(0.08)	(0.18)	(0.09)
Student ELA achievement (Controlling for prior achievement)[c, d]	36,994	0.02	0.04	0.02	0.04
		(0.04)	(0.03)	(0.04)	(0.04)

Notes: *** p<0.01, ** p<0.05, * p<0.1. Standard errors in parenthesis are clustered at the school level. Controlled models use school characteristics—total enrollment, student to teacher ratio, percent of high needs students, percent of ELL students, percent of students with disabilities, and eight school climate survey domain scores from the prior year—and teacher and evaluator characteristics—age, experience, gender, race, and education. We instrument for an evaluator attending any training session and the percent of sessions attended using the treatment indicator.

[a]The sample is subset to teachers that ever met with an evaluator for a post-observation meeting.

[b]These outcomes are created by using the BPS school climate survey, which teachers answered anonymously. Since we cannot link individual teachers to their responses, we only control for school characteristics for these outcomes. Since we cannot link evaluators to teachers' responses, we use treatment to instrument for any evaluator in the school attending any session and all evaluators attending all sessions for the TOT estimates.

[c]Student achievement is measured in grades 3-8 and 10 in mathematics and ELA. We standardize scores at the year, grade, subject level to have a mean of 0 and standard deviation of 1.

[d]Since we include a lagged test score as a control, we exclude from the sample those who did not take the MCAS in the previous year (mostly third graders); this results in a loss of 22% of our sample.

Table A7. *The Effect of Evaluator Training on Teacher and Student Outcomes, Alternative Treatment-Control Contrasts*

		Fall vs Spring			Year 1 vs Spring 2015	
Outcomes	n	Uncontrolled	Controlled	n	Uncontrolled	Controlled
Feedback Quality (PCA, standardized)	2,103	-0.05	-0.06	1,667	-0.09	-0.09
		(0.08)	(0.07)		(0.09)	(0.08)
Ever met to discuss feedback (Binary)	2,210	-0.01	-0.01	1,729	-0.03	-0.02
		(0.04)	(0.03)		(0.05)	(0.05)
Number of observations (Discrete)	2,183	0.34	0.14	1,721	-0.13	-0.17
		(0.37)	(0.38)		(0.34)	(0.35)
Number of discussion meetings (Discrete)	2,210	0.17	0.07	1,729	-0.21	-0.14
		(0.31)	(0.25)		(0.28)	(0.25)
Meeting length (Minutes)	2,210	-1.26	-0.93	1,729	-1.37	-0.94
		(1.02)	(0.92)		(1.37)	(1.28)
Time between observation and meeting (Days)[a]	1,369	0.10	0.57	1,110	0.13	-0.31
		(0.40)	(0.42)		(0.51)	(0.53)
Summative Rating: Curriculum, Planning, and Assessment (Discrete, 1-4)	3,904	0.01	-0.02	3,904	0.00	-0.01
		(0.04)	(0.03)		(0.04)	(0.03)
Summative Rating: Teaching All Students (Discrete, 1-4)	3,904	0.01	-0.03	3,904	0.00	-0.01
		(0.04)	(0.03)		(0.04)	(0.02)
Self-efficacy for classroom management (PCA, standardized)[b]	3,178	0.04	-0.22**	2,528	-0.31*	-0.37**
		(0.14)	(0.11)		(0.18)	(0.17)
Self-efficacy for instructional strategies (PCA, standardized)[b]	3,178	-0.16	-0.14	2,528	-0.68***	-0.50**
		(0.16)	(0.11)		(0.20)	(0.19)
Student math achievement[c]	51,230	-0.02	-0.04	36,779	-0.04	-0.04
		(0.12)	(0.07)		(0.16)	(0.09)
Student math achievement (Controlling for prior achievement)[c, d]	39,996	-0.02	-0.00	28,836	-0.05	-0.03
		(0.12)	(0.03)		(0.17)	(0.04)
Student ELA achievement[c]	51,013	-0.04	-0.05	36,811	-0.09	-0.08
		(0.10)	(0.06)		(0.15)	(0.08)
Student ELA achievement (Controlling for prior achievement)[c, d]	39,457	-0.05	-0.01	28,602	-0.09	-0.07*
		(0.10)	(0.03)		(0.15)	(0.04)

Notes: *** p<0.01, ** p<0.05, * p<0.1. Standard errors are in parenthesis and are clustered at the school level.

This table compares (1) responses from teachers and student achievement from schools that were randomly assigned for fall 2013/2014 to teachers and students from schools that were assigned for spring 2014/2015 and (2) 2015 responses from teachers and student achievement from schools that were randomly assigned in spring 2015 to those randomly assigned in fall 2013 and spring 2014.

Controlled models use school characteristics—total enrollment, student to teacher ratio, percent of high needs students, percent of ELL students, and percent of students with disabilities—and teacher and evaluator characteristics, such as age, experience, gender, race, and education.

[a]The sample is subset to teachers that ever met with an evaluator for a post-observation meeting.

[b]These outcomes are created by using the BPS school climate survey, which teachers answered anonymously. Since we cannot link individual teachers to their responses, we only control for school characteristics for these outcomes.

[c]Student achievement is measured in grades 3-8 and 10 in mathematics and ELA. We standardize scores at the year, grade, subject level to have a mean of 0 and standard deviation of 1.

[d]Since we include a lagged test score as a control, we exclude from the sample those who did not take the MCAS in the previous year (mostly third graders); this results in a loss of 22% of our sample.

Table A8. *Intraclass Correlations of Perceived Evaluation Feedback Quality*

		Proportion of total variance
Panel A. Evaluators		
Evaluator	SD	0.16
	SE	0.02
Error	SD	0.84
Panel B. Schools and Evaluators		
School	SD	0.04
	SE	0.02
Evaluator	SD	0.13
	SE	0.02
Error	SD	0.83
Panel C. Schools, Evaluators, and Teachers		
School	SD	0.04
	SE	0.02
Evaluator	SD	0.13
	SE	0.02
Teacher	SD	0.57
	SE	0.02
Error	SD	0.27

Notes: We use pooled data from SY 2013-14 and SY 2014-15. SD = Standard Deviation. SE = Standard Error.

Table A9. *The Relationship Between School Characteristics and Perceived Evaluation Feedback Quality*

	Preferred model with no summative rating	Preferred Model	Teachers rated *Proficient* only	School fixed effects	Evaluator fixed effects	Weighted by teacher survey response rate	Weighted by propensity to take survey
Collegial work environment	-0.084*	-0.048	-0.081	-0.157**	-0.028	-0.047	-0.051
	(0.046)	(0.045)	(0.053)	(0.072)	(0.075)	(0.050)	(0.043)
School leadership quality	0.160***	0.165***	0.179***	0.024	0.029	0.173***	0.163***
	(0.035)	(0.032)	(0.037)	(0.049)	(0.048)	(0.034)	(0.031)
Parent and student engagement	-0.110	-0.116	-0.119	-0.079	-0.162	-0.147*	-0.104
	(0.078)	(0.077)	(0.086)	(0.128)	(0.132)	(0.080)	(0.077)
Collective teacher efficacy	0.011	-0.045	-0.018	0.064	0.083	-0.063	-0.040
	(0.057)	(0.055)	(0.063)	(0.084)	(0.084)	(0.059)	(0.053)
Self-efficacy for classroom management	0.025	0.018	-0.008	0.009	0.011	0.026	0.015
	(0.053)	(0.050)	(0.054)	(0.089)	(0.074)	(0.052)	(0.048)
Teacher influence over classroom decision-making	-0.055*	-0.072**	-0.086**	0.004	0.021	-0.072**	-0.070**
	(0.031)	(0.029)	(0.035)	(0.078)	(0.062)	(0.031)	(0.029)
Self-efficacy for instructional strategies	0.051	0.077**	0.061	0.014	-0.048	0.085**	0.077**
	(0.034)	(0.032)	(0.037)	(0.047)	(0.036)	(0.036)	(0.031)
Relationship with students and parents	0.078*	0.069	0.075	-0.052	-0.094	0.082*	0.064
	(0.047)	(0.047)	(0.053)	(0.077)	(0.068)	(0.049)	(0.046)
Enrollment	-0.000	-0.000	-0.000	-0.001	0.001***	-0.000	-0.000
	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)
Student to teacher ratio	0.013	0.018	0.007	0.018	0.018	0.020	0.016
	(0.011)	(0.012)	(0.014)	(0.021)	(0.014)	(0.013)	(0.011)
Percent of high needs students	0.002	0.001	-0.003	-0.007	-0.007	0.001	0.002
	(0.004)	(0.004)	(0.005)	(0.006)	(0.006)	(0.004)	(0.004)
Percent of ELL students	-0.004*	-0.002	-0.003	-0.001	0.001	-0.002	-0.003
	(0.002)	(0.002)	(0.003)	(0.010)	(0.006)	(0.002)	(0.002)
Percent of students with disabilities	0.002	0.003*	0.003	-0.002	0.010*	0.003	0.003
	(0.002)	(0.002)	(0.002)	(0.008)	(0.006)	(0.002)	(0.002)
Percent of students African-American	-0.006**	-0.006**	-0.003	-0.014**	-0.008**	-0.006**	-0.006***
	(0.002)	(0.002)	(0.003)	(0.007)	(0.004)	(0.003)	(0.002)
Percent of students Asian	-0.009**	-0.009**	-0.004	-0.006	-0.026***	-0.009**	-0.009**
	(0.004)	(0.004)	(0.005)	(0.024)	(0.009)	(0.004)	(0.004)

Percent of students Hispanic	-0.003	-0.003	0.003	-0.000	-0.018***	-0.004	-0.003
	(0.003)	(0.003)	(0.004)	(0.011)	(0.005)	(0.004)	(0.003)
Percent of students Other	0.009	-0.001	0.023	0.028	-0.005	0.002	-0.000
	(0.024)	(0.025)	(0.028)	(0.053)	(0.041)	(0.028)	(0.024)
Middle school	0.036	0.031	-0.027		-0.505**	0.013	0.040
	(0.072)	(0.069)	(0.082)		(0.194)	(0.075)	(0.067)
High school	0.163	0.106	0.097		-0.470	0.070	0.106
	(0.114)	(0.109)	(0.124)		(0.459)	(0.119)	(0.105)
Control for summative ratings	N	Y	Y	Y	Y	Y	Y
Survey response weights	N	N	N	N	N	Y	N
School fixed effects	N	N	N	Y	N	N	N
Evaluator fixed effects	N	N	N	N	Y	N	N
n	4,103	4,103	3,092	4,103	4,103	4,103	4,103

Notes: *** p<0.01, ** p<0.05, * p<0.1. Standard errors are in parenthesis.

Models use pooled data from SY 2013-14 and SY 2014-15 and estimate the relationship between evaluation feedback quality and teacher, evaluator, and school characteristics. This table contains the school characteristics estimates that are not shown in Table 5 – for the estimates for teacher and evaluator characteristics refer to Table 6. Measures of school climate are based on teachers' responses to the BPS Climate Survey and are aggregated up to the school level. All models contain fixed effects for school year and have standard errors clustered at the school level.

Appendix A Figures

Conversation Observation Note Taking Guide

Teacher: **Observer:** **Supervisor:**

Conversation Protocol	Principal's Questions/Prompts/Suggestions	Teacher's Questions/Prompts/Insights
• Ask the teacher to summarize the impressions of the lesson		
• Ask the teacher to recall data to support those impressions		
• Analyze the observation evidence together		
• Help the teacher synthesize the evidence and decide next steps		
• Reflect on the process and propose refinements		

Appendix Figure A1. Example of observation and feedback tool

Appendix B

The validity of our perceptions of feedback quality measure. To examine the construct validity of our measure of perceived evaluation feedback quality, we explore its relationship with other measures theoretically and empirically linked to high-quality evaluation and feedback. Prior studies have found that teachers find feedback useful when it is evidenced-based, timely, and in-depth (Cherasaro et al., 2016). Teachers that are observed more often and that have more immediate, frequent, and longer meetings with their evaluators are likely to report receiving higher-quality feedback. We find this to be the case – our measure of perceived feedback quality is positively correlated with the number of times a teacher is observed by their evaluator ($r = 0.28$), the number of discussion meetings teachers have with their evaluator ($r = 0.25$), and the length of post-observation discussion meetings ($r = 0.10$), while being negatively correlated with the time between being observed and having a discussion meeting ($r = -0.08$).

We explore the predictive validity of perceived feedback quality by examining the relationship between our measure and changes in teachers' performance. We measure teacher performance in three ways: 1) the BPS summative rating, 2) the average of the four BPS subdomain ratings, and student achievement. We regress the change in teachers' evaluation ratings between the current year and the prior year on the current year's perceived feedback quality, controlling for evaluator and school characteristics. We also regress gain scores in student math and ELA achievement on perceived feedback quality, controlling for student and school characteristics. As reported in Appendix Table B1, we find that perceptions of higher-quality feedback are associated with gains in teacher performance as measured by their evaluation ratings. We find a one SD increase in perceived feedback quality is associated with a 0.07 point increase in a teacher's summative rating score on the 4-point scale, and a similar magnitude for the subdomain score (0.05). However, we find no relationship between perceived

feedback quality and student achievement gains. These mixed results suggest that higher-quality evaluation feedback may support professional growth as measured broadly by the evaluation process, but does not necessarily contribute to changes in instruction that improve student performance on standardized tests. It is likely that receiving high-quality feedback and perceiving it as such is a necessary, but not always a sufficient condition for evaluation feedback to improve teacher performance.

Table B1. *The Relationship Between Perceived Evaluation Feedback Quality and Gains in Teacher Effectiveness and Student Achievement*

	Gain in BPS Rating[a]		Gain in Math Score[b]		Gain in ELA Score[b]	
	(1)	(2)	(3)	(4)	(5)	(6)
Perceived Evaluation Feedback Quality	0.068***	0.072***	-0.006	-0.001	0.001	0.006
	(0.010)	(0.010)	(0.010)	(0.009)	(0.010)	(0.010)
School fixed effects	N	Y	N	Y	N	Y
n	3,579	3,579	42,372	42,372	41,814	41,814

Notes: *** p<0.01, ** p<0.05, * p<0.1. Standard errors are in parenthesis.

The models use pooled data from SY 2013-14 and SY 2014-15 and estimate the relationship between changes in teacher effectiveness and student achievement over the previous year on teachers' perceived evaluation feedback quality. The outcome for models (1) and (2) is the gain in a teacher's BPS overall summative rating over the previous year, for models (3) and (4) it is the gain in a student's MCAS math score, and for models (5) and (6) it is the gain in a student's MCAS ELA score. The second column for each outcome uses school fixed effects. All models include fixed effects for the school year and standard errors clustered at the school level.

[a]For columns 1-2, we control for evaluator and school characteristics. These include evaluator characteristics such as age, tenure at school, gender, and race/ethnicity. For school characteristics, we include total enrollment, student-to-teacher ratio, percent of high needs students, percent of students by race, percent of ELL students, percent of students with disabilities, and eight one year lagged domains from the BPS school climate survey.

[b]For columns 3-6, we control for student race, gender, special education status, eligibility for free or reduced price-lunch, and grade level. We include the following school level controls: total enrollment, student-to-teacher ratio, percent of high needs students, percent of students by race, percent of ELL students, and the percent of students with disabilities.