ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 19-62

In Search of High-Quality Evaluation Feedback: An Administrator Training Field Experiment

Matthew A. Kraft
Brown University

Alvin Christian
Brown University

Starting in 2011, Boston Public Schools (BPS) implemented major reforms to its teacher evaluation system with a focus on promoting teacher development. We administered independent district-wide surveys in 2014 and 2015 to capture BPS teachers' perceptions of the evaluation feedback they receive. Teachers generally reported that evaluators were fair and accurate, but that they struggled to provide high-quality feedback. We conduct a randomized controlled trial to evaluate the district's efforts to improve this feedback through an intensive training program for evaluators. We find little evidence the program affected evaluators' feedback, teacher retention, or student achievement. Our results suggest that improving the quality of evaluation feedback may require more fundamental changes to the design and implementation of teacher evaluation systems.

VERSION: May 2019

Suggested citation: Kraft, M.A. & Christian, A. (2019). In Search of High-Quality Evaluation Feedback: An Administrator Training Field Experiment (EdWorkingPaper No.19-62). Retrieved from Annenberg Institute at Brown University: http://edworkingpapers.com/ai19-62

In Search of High-Quality Evaluation Feedback: An Administrator Training Field Experiment

Matthew A. Kraft
Brown University

Alvin Christian
Brown University

March 2019

Abstract

Starting in 2011, Boston Public Schools (BPS) implemented major reforms to its teacher evaluation system with a focus on promoting teacher development. We administered independent district-wide surveys in 2014 and 2015 to capture BPS teachers' perceptions of the evaluation feedback they receive. Teachers generally reported that evaluators were fair and accurate, but that they struggled to provide high-quality feedback. We conduct a randomized controlled trial to evaluate the district's efforts to improve this feedback through an intensive training program for evaluators. We find little evidence the program affected evaluators' feedback, teacher retention, or student achievement. Our results suggest that improving the quality of evaluation feedback may require more fundamental changes to the design and implementation of teacher evaluation systems.

In Search of High-Quality Evaluation 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, Wingert, Conant, & Register, 2010) and promoting professional development through classroom observations and feedback (Almy, 2011; Curtis & Wiener, 2012; Papay, 2012). Most states and districts have emphasized the latter goal of improving teachers' instruction (Center on Great Teachers and Leaders, 2014). Several years into this national policy experiment, we still know little about the quality of feedback teachers receive as part of these new high-stakes evaluation systems.

A growing body of evidence affirms the potential of frequent teacher feedback to drive instructional improvement and student achievement. Extensive feedback is a core organizational practice of both effective urban charter schools (Angrist, Pathak, & Walters, 2013; Dobbie & Fryer, 2013) and high-performing traditional public schools (Reinhorn, Johnson, & Simon, 2017). Randomized controlled trials demonstrate that teacher coaching programs centered around individualized, context-specific feedback have sizable effects on teachers' classroom practices and students' academic performance (Kraft, Blazar, & Hogan, 2018). Several studies have also found that observation and feedback cycles that leverage data from evaluation systems in low-stakes settings can increase student achievement (Garet et al., 2017; Papay, Taylor, Tyler, & Laski, 2016; Steinberg & Sartain, 2015).

Although most state plans have emphasized using evaluation reforms to drive teacher development, many of these systems have been designed and implemented in ways that prioritize teacher accountability. Investments in new systems have skewed heavily towards efforts to

develop and use new evaluation measures to better differentiate teachers' performance (Institute for Education Science, 2014). Most new evaluation systems require only two formal and two informal observations and do not consistently mandate post-observation conferences (Steinberg & Donaldson, 2016). The vast majority of districts have simply added the responsibility of conducting regular teacher observations to administrators' existing tasks (Kraft & Gilmour, 2016; Neumersiki et al., 2018). Efforts to support administrators have focused on familiarizing them with new observation instruments and improving their reliability as evaluators, rather than on their ability to provide high-quality instructional feedback (Herlihy et al., 2014).

In this paper, we examine teachers' perceptions of the feedback they receive as part of Boston Public Schools' (BPS) new teacher evaluation system and evaluate the district's efforts to strengthen the quality of this feedback. In the 2011-12 academic year, BPS implemented major reforms to its evaluation system, with a focus on using the evaluation process as a tool for teacher development. The following year, BPS convened a group of experienced administrators to develop and pilot a multi-day evaluator training program intended to improve the quality of feedback administrators provided to teachers.

We evaluate the implementation and effects of this intensive 15-hour evaluator training series by exploiting the staggered rollout of the program across two academic years. Working with BPS, we randomly assigned school-based evaluation teams to attend the training program in one of the four semesters across the 2013-14 and 2014-15 school years. This semester-specific randomization design allows us to test a range of treatment-control contrasts. We also administered an independent, district-wide survey to teachers to learn about their experiences with the evaluation process at the end of these two school years. Teachers' responses provide a

range of outcomes for the randomized controlled trial (RCT) and allow us to extend prior research on teachers' experiences with evaluation feedback.[1]

Analyzing the quality of teacher feedback generated as part of high-stakes evaluation systems, the correlates of high-quality feedback, and the potential to improve this feedback advances our understanding of a key mechanism in the theory of action for how evaluation reforms were intended to promote teacher development. We examine teachers' *perceptions* of the feedback they receive because both pedagogical theory and prior empirical research suggest that teachers are unlikely to respond to feedback in constructive ways if they don't believe it to be of high quality (Garubo & Rothstein, 1998; Feeney, 2007; Cherasaro et al., 2016). We find that BPS teachers generally thought that evaluators were fair and accurate raters, but viewed the quality of feedback they received less favorably. Ultimately, just over a quarter of teachers felt that their instruction improved because of this feedback.

We next explore what teacher, evaluator, and school characteristics are correlated with feedback that teachers perceived to be of high quality. Holding constant teachers' overall evaluation ratings, we find that less-experienced teachers are more likely to rate the feedback they receive as higher-quality, and that evaluators with longer tenures in their current schools are perceived to provide better feedback. We also find that teachers of color who are evaluated by administrators of the same race report receiving substantially higher-quality feedback than others. These results are consistent with prior research that document the positive benefits of teacher and student racial congruence for student outcomes (Dee, 2004, 2007; Egalite, Kisida, & Winters, 2015; Lindsay & Hart, 2017; Egalite & Kisida, 2018; Holt & Gershenson, 2015; Gershenson, Hart, Hyman, Lindsay, & Papageorge, 2018).

[1] See Donaldson, 2012; Donaldson, 2016; Neumerski et al., 2018; Jiang, Sporte, & Luppescu 2015; Donaldson et al., 2014; Cherasaro, Brodersen, Reale, & Yanoski, 2016; Firestone, Nordin, Shcherbakov, Kirova, & Blitz, 2014.

The evaluator training program we evaluate was reasonably well attended by administrators and implemented successfully. Administrators reported high levels of satisfaction with the training and increases on a range of self-evaluated skills from baseline to end-of-course surveys. However, we find little evidence to suggest that the intensive training program was successful at improving the perceived quality of evaluation feedback or affected the frequency or duration of observation and feedback cycles in the short or medium-term. Together, our descriptive and causal evidence shed new light on the potential and limitations of promoting professional development through the teacher evaluation process. These findings can inform states' and districts' ongoing efforts to redesign their teacher evaluation systems under the increased flexibility provided by the Every Student Succeeds Act (ESSA).

Background

Teacher Evaluation Reforms and Teacher Effectiveness

Evidence of the effect of high-stakes teacher evaluation reforms on teacher performance and student achievement is decidedly mixed (Jackson & Cowan, 2018). There are several potential channels through which evaluation might increase teacher effectiveness. One pathway is through changing the composition of the teacher workforce by tying high-stakes personnel decisions to evaluation ratings. Studies have shown, for example, that the introduction of high-stakes evaluation systems in several urban districts has increased the voluntary attrition of low-performing teachers (Dee & Wyckoff, 2015; Adnot, Dee, Katz & Wyckoff, 2017; Sartain & Steinberg, 2016; Cullen, Koedel & Parsons 2016). Emerging evidence also shows that new high-stakes evaluation systems increased the probability new teaching candidates attended more

competitive undergraduate institutions, while also decreasing the overall supply of teaching candidates (Kraft, Brunner, Dougherty & Schwegman, 2019).

A second pathway is through improving the performance of current teachers. Such improvements might reflect professional growth on the job due to evaluation feedback and/or increased effort incentivized by dismissal threats or merit pay. Several quasi-experimental and experimental studies in large urban districts point to the potential of evaluation systems to serve as engines for teacher professional growth. Taylor and Tyler (2012) analyzed an evaluation system in Cincinnati Public Schools centered around classroom observations and feedback. Using data from 2003-04 to 2009-10, they found that being evaluated improved teachers' ability to raise student math achievement, but had no effect on reading achievement. Steinberg and Sartain (2015) found that a pilot implementation of the new teacher evaluation system in Chicago Public Schools produced significant improvements in reading achievement and positive, but imprecisely estimated, effects in math in the first year. However, the authors found no effect in either subject among the cohort of schools who adopted the system in the second year. Dee and Wyckoff (2015) found that the high-powered incentives in the District of Columbia Public Schools' IMPACT program resulted in teachers at risk of dismissal and close to receiving large bonuses to improve their performance the following year.

However, we know much less about whether the high-stakes evaluation systems recently implemented in most states are successful at promoting teacher professional growth. At a minimum, these systems have introduced rigorous classroom observation rubrics which provide a common language for discussing high-quality instruction. Research also suggests principals primarily rely on data from these classroom observations to inform their personnel decisions (Goldring et al., 2015). At the same time, observation ratings may be limited in their utility for

informing professional development and feedback because they can be inflated, biased, and unreliable. Despite reformers' efforts, nearly all teachers continue to be rated satisfactory or higher (Kraft & Gilmour, 2017) and several studies have documented that evaluators' observation scores are influenced by the student composition of their classrooms (Whitehurst, Chingos, & Lindquist, 2014; Steinberg & Garret, 2016; Campbell & Ronfeldt, 2018). Observation ratings on specific sub-domains, which have the most potential to inform targeted feedback, also appear to be relatively unreliable (Garet et al., 2017).

Large-scale studies of evaluation reforms point to significant challenges of taking new evaluation to scale across diverse contexts. Garet and his colleagues (2017) found that introducing frequent observation and feedback cycles as part of a low-stakes evaluation system had mixed results across eight districts. Teacher practice improved as judged by scores on the CLASS rubric, but not on the Danielson rubric. Student achievement in math also improved very modestly in the first year, but the increased feedback had no statistically significant effect on math achievement in the second year or reading achievement in either year. Researchers at RAND evaluated the efforts of three school districts and four charter management organizations to implement sweeping evaluation and human capital reforms supported by large grants (Stecher et al., 2018). They found that attempts to differentiate professional development based on individual performance were seen as ineffective by teachers, and only one of six sites implemented more than one observer per teacher. Ultimately, these reforms failed to drive changes in instruction or achievement.

Teacher Evaluation Feedback on The Ground

The mixed effects of evaluation reforms are likely the product of varying implementation quality across states and districts. The success of evaluation reforms requires navigating a long

mediational chain of steps to impact teacher and student outcomes. Studying evaluation reforms in Connecticut, Donaldson (2012, 2016) found that teachers' valued the observation and feedback process, but felt they did not receive enough guidance about crafting goals or feedback about their instruction. Reinhorn, Johnson, and Simon (2017) found that high-performing schools in Massachusetts serving large populations of students from low-income families prioritized using their new high-stakes evaluation system as a development tool. A majority of teachers at these schools reported being observed and receiving feedback *at least* 5 to 10 times a year, and nearly every teacher described the feedback as supporting their instructional improvement. Studies also have shown that teachers at schools with fewer resources and more pressing student needs receive less frequent and lower-quality feedback (Donaldson, Woulfin, LeChasseur, & Cobb, 2016).

The success of evaluation reforms depends heavily on the skills and capacity of evaluators. Administrators employ substantial discretion in how they implement evaluation systems, which can enhance or undermine the quantity and quality of feedback teachers receive (Donaldson & Woulfin, 2018). Many new evaluation systems require administrators to spend substantially more time observing, scoring, and providing feedback to teachers. These reforms have frustrated some administrators that believe these additional duties prevent them from managing their traditional responsibilities as school leaders, such as meeting with parents and students (Neumerski et al., 2018). Moreover, reforms created expectations for administrators to provide feedback for teachers in grades and subjects where they have limited prior experience (Kraft & Gilmour, 2016). In Chicago, researchers found that administrators largely dominated post-observation conversations and rarely asked mid- and high-level questions that pushed teachers to reflect on their practices (Sartain, Stoelinga, & Brown, 2011).

Several reports commissioned by federal and state education agencies have examined teacher's experiences during the pilot phases of new evaluation systems using data from anonymous teacher surveys (Donaldson et al., 2014; Cherasaro et al., 2016; Firestone et al., 2014). These studies of new evaluation systems in Connecticut, New Jersey, and two unidentified mid-western states found consistent evidence that teachers generally viewed the evaluation process as accurate and credible, but had much more mixed responses about the timeliness and usefulness of the feedback they received. Most relevant to our descriptive analyses, Jiang, Sport, and Luppescu (2015) analyzed teachers' perspectives on evaluation reforms in Chicago Public Schools (CPS) in the first two years of implementation. Similar to the reports describe above, they found evidence that teachers judged their evaluators as fair and accurate. However, they also found that the vast majority of CPS teachers reported that the feedback they received included specific and actionable feedback from evaluators. We extend this literature by examining teachers' experiences in Boston Public Schools during the second and third years of adopting a new high-stakes evaluation system.

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 in partnership with their principals, collect evidence of their progress towards these goals, are observed by principals, and participate in a formative and summative evaluation process. Evaluations are based on a rubric, developed by the state and adapted by BPS, that is comprised

of four broad domains: 1) Curriculum, Planning, and Assessment, 2) Teaching All Students, 3) Family and Community Engagement, and 4) Professional Culture.

Principals and select members of school administrative teams (e.g., assistant principals, directors of instructors) serve as evaluators. Evaluators conduct one to four formal unannounced observations of each teacher throughout the year and provide formal written feedback depending on a teacher's prior evaluation rating. In addition, evaluators are encouraged to conduct frequent informal observations lasting 15-20 minutes and hold face-to-face post-observation conversations with teachers. Unlike many evaluation systems that apply a weighted formula and pre-established score thresholds to determine teachers' overall summative rating (Steinberg & Kraft, 2017), BPS administrators are individually responsible for assigning overall and domain-specific summative ratings using a four-point scale ranging from *Unsatisfactory* to *Exemplary*. Evaluators' holistic ratings are based on classroom observations and evidence submitted by teachers documenting their progress towards professional practice and student learning goals. Teachers rated in the top two categories continue this cycle of self-directed growth, while those in the lower rating categories are placed on more structured evaluation plans, which, after several repeated low evaluations, can result in dismissal.

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 spring of the 2012-13 school year. The training sequence emphasized the importance of 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 curriculum was grounded in theories of adult learning and provided evaluators with practical strategies for conducting classroom observations and providing feedback. For example, 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 program was designed around the ideas that in-person conversations can strengthen relationships and reduce the chance for misunderstandings, and that clear, specific, and actionable feedback is essential for supporting ongoing professional growth. The program did not focus on developing administrators' specific content knowledge, but instead emphasized teaching evaluators more general techniques to develop their relationships with teachers and produce effective feedback. For example, trainers taught evaluators to base their feedback on observable evidence and to use coaching language instead of evaluative language to describe observations. The training also provided evaluators with specific language, techniques, and rubrics to conduct post-observation discussion meetings. These tools were intended to encourage teachers to be more receptive to feedback, reflect in-depth about their practices, and push them grow.

There are a few key practices that differentiate this training from traditional 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 strong guiding philosophies and theories, but also included practical strategies throughout every session; 3) participants completed homework between each session

to try out 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, though some were on weekends. Research team members attended training sessions to track attendance, take field notes, and administer surveys.

Methods

Data and Measures

Schools. Our sample of 123 BPS schools includes traditional public schools, charter schools, and pilot schools.[2] As we show in Table 1, BPS enrolls about 60,000 students, the vast majority of whom are students of color and considered high needs (84%).[3] The district also serves a sizeable population of ELL students (31%) and students with disabilities (21%).

Teachers. Between 2013-2015, BPS employed a total of 4,805 teachers within the 123 schools in our sample. 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. Teacher experience varied widely across the district.

Independent teacher survey. We administered an independent, confidential, but individually identifiable, survey to teachers to capture their views on the evaluation process at the end of the 2013-14 and 2014-15 school years, the second and third years of implementation for the new teacher evaluation system. The survey consisted of 29 items measured on a five-point Likert scale. Questions examined teachers' perceptions about evaluators' communication,

[2] Pilot schools are semiautonomous district schools that have autonomy over budgeting, staffing, governance, curriculum/assessment, and the school calendar.
[3] 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.

fairness, utility, feedback, and relationship quality. The survey also captured more concrete evaluation implementation information, such as the number of unannounced and announced observation visits evaluators made, the number of post-observation meetings evaluators and teachers had, and the length of these meetings.

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.[4] 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%.

Our efforts resulted in responses from 56% of teachers in the 2013-14 school year and 60% in the 2014-15 school year. These response rates are notably higher than prior research efforts to collect independent district-wide teacher surveys (e.g. Ronfeldt, Farmer, McQueen, & Grissom, 2015) and comparable favorably to the federally-funded 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 important ways as shown in Table 2. For example, in 2013-14 teachers who completed the post-evaluation survey were more likely to be female (77% vs.

[4] Teacher leader is a formal position in BPS held by at least one classroom teacher in each school to act as a liaison between the district and schools.

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. For some teachers, feedback came only in late May as a formal written evaluation. The end-of-year survey administration also meant that most teachers had received their summative evaluation rating prior to completing the survey. 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%). These patterns persisted in 2014-15, but were somewhat less pronounced.

Although these differential response rates may limit the generalizability of our findings, our data include more than half of all teachers in BPS – a sizable and diverse sample in its own right. A second possible concern might be that teachers' survey responses were influenced by knowing their summative evaluation ratings. We take several steps to guard against this potential bias in our correlational analyses. As described below, we control for teachers' summative evaluation ratings in all regressions, and we confirm our results are robust to limiting our sample to only those teachers who received the exact same *Proficient* summative rating. We also find no evidence that teachers' response rates were related to the timing of when they were randomly assigned to attend the training series. Thus, the differential response rates described above do not pose a threat to the causal estimates from our RCT.

Teachers' perceptions of evaluation feedback quality. We create a latent measure of teachers' perceptions about the quality of evaluation feedback they received using eight items from our independent teacher survey. Constructing a scale based on multiple items serves to

minimize measurement error; together the 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 A1 for the full set of 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. We also create a simple alternate measure of perceptions of evaluation feedback quality for descriptive purposes by assigning values of 1-5 to the Likert scale responses and averaging across items.

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. Specifically, we regress gain scores in teachers' overall evaluation ratings between the current year and the prior

year on 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.[5] As reported in Appendix Table A2, we find that perceptions of higher-quality feedback are associated with gains in teacher performance as measured by their overall evaluation rating. We find a one SD increase in perceived feedback quality is associated with a 0.07 point increase in a teacher's rating score on the 4-point scale. However, we find no relationship between perceived feedback quality and student achievement gains. These mixed results suggest that higher-quality evaluation feedback supports 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.

BPS teacher survey. BPS administers an anonymous annual school climate survey to teachers which asks about the school leadership and work environment, classroom instruction, management, autonomy, and engagement and relationships with parents and students. The survey consists of 62 questions measured on a four-point Likert scale. We use these data to create three outcome measures: self-efficacy for instructional strategies, self-efficacy for classroom management, and quality of school leadership. Self-efficacy for instructional strategies consists of three items ($\alpha = 0.85$), self-efficacy for classroom management consists of six items ($\alpha = 0.83$), and quality of school leadership consists of fifteen items ($\alpha = 0.97$). For each domain, we predict the first principal component from a factor analysis and standardize it to have a mean of 0 and SD of 1.

[5] See the Notes section of Appendix Table A2 for a full description of the covariates included in these models.

Evaluators. 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 3. Similar to teachers, the majority of evaluators were female (70%). Notably, a larger percentage of evaluators were evaluators of color compared to teachers (52% vs 39%). The typical evaluator had been in their current administrative position just over three years, though some had almost 30 years of tenure at their schools.

Independent evaluator surveys. We administered surveys to evaluators both at the beginning and end of the training series to gather information on three domains. The survey consisted of fifteen questions on a nine-point Likert scale. Questions covered a range of topics including asking evaluators' opinions about the evaluation system, the quality of the training, and their own ability to provide constructive feedback. 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.

Correlational Analyses

We begin by exploring the relationship between perceived evaluation feedback quality and a range of predictors using Ordinary Least Squares (OLS) regression. We model perceived evaluation feedback quality for teacher *i* at school *s* in year *t* as follows*:*

$$Evaluation\ Feedback_{ist} = \alpha + \beta X_{ist} + \gamma_t + \varepsilon_{ist} \quad (1)$$

Here *X* represents a vector of teacher, evaluator, and school characteristics. These include teacher and evaluator characteristics such as age, experience, gender, race/ethnicity, and education level.[6] We also include indicators for the overall summative rating category teachers

[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.

received as part of the evaluation process. For school characteristics, we include total enrollment, student-to-teacher ratio, percent of students by race, percent of high-needs students, percent of English language learners, percent of students with disabilities, and measures of the eight school climate survey domain scores from the prior year. We include fixed effects for year, γ, and cluster standard errors at the school level.

Randomization Design and Analyses

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 in a given semester (fall 2013, spring 2014, fall 2014, or spring 2015). We grouped eligible schools into six blocks based on school size (small vs. large) and type (elementary, middle, and high) and then randomized within school size-type blocks. School teams could then choose to attend one of three series of training sessions offered at different times each semester. In Table 1, we show that schools were balanced on observable characteristics across all four randomization groups.

The semester-specific random design allows us test multiple treatment-control contrasts as well as to examine heterogeneity by the timing of training. Our primary treatment-control contrast identifies the effect of being randomly assigned to attend the training program in the first year of the program on outcomes at the end of that year. Here, evaluators 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. We estimate the effects of being randomly assigned to attend the evaluator training program during the 2013-14 school year on a range of outcomes using the following OLS model:

$$Y_{ist} = \alpha + \beta Treat_{st} + \delta X_{ist} + \gamma_t + \pi_b + \varepsilon_{ist} \qquad (2)$$

The outcome Y_{ist} represents a teacher or student outcome such as the perceived quality of evaluation feedback or student achievement. $Treat_{st}$ is an indicator for a given definition of treatment. For each outcome, we present both results from baseline models without controls as well as conditional estimates from models that include controls to increase precision and test the sensitivity of our findings. For teacher-specific outcomes, we control for teacher, evaluator, and school characteristics included in X as described above. For student achievement 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 fixed effects for year, γ, and school size-type blocks, π, to account for the stratified randomization process. We again cluster standard errors at the school level.

We also estimate two additional treatment-control contrasts using equation (2). We examine the effect of being randomly assigned to attend the training in the fall verses the spring semester to assess if receiving training earlier in the year had a larger effect on evaluators' practices. 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. 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 the 2013-14 school year and the control group as evaluators randomized to attend during spring 2015. This third treatment-control contrast serves as a lower-bound estimate of any medium-term effects given the control group was also treated just before the outcomes we examine at the end of the second year of the training program (2014-15).

Findings

Assessing Teachers' Performance

BPS evaluation reforms were intended to create a system where all teachers would receive regular, accurate evaluations as well as high-quality feedback about how they could improve their performance on the job. According to teachers, BPS evaluators were successful at evaluating teachers regularly and fairly. 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. Evaluators also produced ratings that reflected limited, but still meaningful variation in performance across teachers. Across both years, approximately 6% of teachers were rated *Unsatisfactory* or *Needs improvement*, the bottom two ratings, while 18% of teachers were rated *Exemplary*, the highest possible rating. This distribution of ratings reflects a strong skew towards higher ratings, but also greater differentiation than most teacher evaluation systems (Kraft & Gilmour, 2017).

Teachers generally believed that evaluators were fair and accurate, and they felt 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.

Providing Performance Feedback to Teachers

Although teachers generally thought that evaluators were fair and accurate raters, teachers had far less favorable views about the quality of feedback they received as part of the

evaluation process. In Figure 1, 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 of teachers felt that their feedback was useful or actionable. Ultimately, just over a quarter of teachers felt that their instruction improved because of this feedback.

Teachers' general dissatisfaction with feedback was compounded by evaluators struggling to find time to meet and provide feedback. Evaluators observed teachers, on average, almost six times a year but only met with each teacher an average of two times to provide feedback. Over one-third of teachers reported *never* meeting with an evaluator to have a post-observation feedback discussion. In Figure 2, 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 providing feedback to each teacher. Furthermore, only about half of the teachers surveyed said that their evaluator was effective at communicating feedback. Evaluators rarely pushed teachers to be active participants during feedback conversations – only about one-third of teachers said that their evaluator would ask questions that allowed them to reflect in-depth about their teaching practices.

A number of challenges likely contributed to the perceived lack of consistent, high-quality feedback from evaluators. BPS administrative data show that the typical evaluator assessed about a dozen teachers a year. Approximately 17% of evaluators in the second year of the new evaluation system and 10% in the third year evaluated 20 or more teachers. This high teacher-to-evaluator ratio likely constrained evaluators' ability to provide more extensive personalized feedback to teachers.

The design of the district evaluation system also directed evaluators' focus towards formal written feedback rather than in-person 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 a majority 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; only 46% of teachers felt that the evaluation system's primary purpose was to help teachers improve.

Variability in the Perceived Quality of Evaluation Feedback

We find that the simple descriptive statistics presented above mask considerable heterogeneity in the perceived quality of evaluation feedback. A simple variance decomposition of perceived feedback quality across and within evaluators suggests that some evaluators are substantially more effective at providing feedback than others. We find that 16% of the variance of our perceived feedback quality measure is between evaluators. In Figure 3, 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%) have average perceived feedback quality ratings that are statistically significantly above the mean, and 37 (11%) have ratings that are statistically significantly below the mean. The 26% of evaluators with average ratings that are significantly different from the mean is far larger than what would be expected by chance given a null hypothesis of there being no difference in perceived feedback quality across evaluators (5%).

In Figure 4, we display the distribution of perceived evaluation feedback quality at the teacher level by overall evaluation ratings. Two important patterns emerge: 1) less effective teachers, as measured by their overall summative rating, report receiving lower-quality feedback, and 2) there is considerable variability in perceived feedback quality even among teachers with the same evaluation rating. The first pattern could be explained by: a) the fact that evaluators struggled to communicate more negative feedback to lower-performing teachers, b) reporting bias where teachers who received higher ratings judged the feedback they received as higher-quality, or both. The considerable variability in perceived feedback quality even among teachers with the same evaluation rating, however, illustrates that teachers' overall summative ratings are not the driving factor in their assessments of the quality of feedback they report receiving.

Correlates of High-Quality Evaluation Feedback

Given the wide variation in teachers' experiences with evaluation feedback, we seek to understand the relationship between teacher, evaluator, and school characteristics and how teachers perceive the quality of feedback they receive. Positive relationships between characteristics and perceived feedback quality may imply certain characteristics are associated with receiving objectively higher-quality feedback, but they may also imply that teachers with particular characteristics are more satisfied with the feedback they receive regardless of the actual quality of feedback. We find that school-level characteristics are, overall, only weakly associated with perceived evaluation feedback quality (see Appendix Table A3), so we focus our discussion on teacher and evaluator characteristics.

Teacher characteristics. Our estimates in Table 4 indicate that among teacher characteristics, experience and race are the strongest predictors of perceived feedback quality. In column 1, our preferred model specification, we show that less experienced teachers report

receiving substantially higher-quality feedback than their more experienced peers. Compared to teachers with 0-2 years of teaching experience, teachers with 9+ years of experience report that their feedback quality is 0.33 SD lower. Race also appears to be a strong predictor of teachers' perceptions of evaluation feedback quality; African-American, Asian, and Hispanic teachers report receiving higher-quality feedback than white teachers, even when controlling for their evaluation rating and evaluator and school characteristics. Being a teacher of color is associated with 0.13-0.25 SD higher reported evaluation feedback quality relative to white teachers.

Evaluator characteristics. Among evaluator characteristics, we find that tenure at a school and race are both important predictors. 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 0.16 SD higher. We find that patterns by race run in the opposite direction for evaluators than for teachers. Compared to white evaluators, the perceived quality of evaluation feedback is lower for evaluators of color. 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 concerning questions about whether some teachers are less receptive to feedback, or more critical of it, when it comes from evaluators of color.

Racial congruence. Given the growing literature documenting the importance of racial congruence between teachers and students for student outcomes and the sizable samples of teachers of color (39%) and evaluators of color (52%) employed by BPS, we seek to understand how evaluator and teacher racial congruence is related to perceived feedback quality. We find that racial congruence in teacher and evaluator pairs is important in explaining 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 0.29 SD higher than racially incongruent pairs. We find positive estimates of similar magnitudes for racial congruence among Hispanics and Asians of 0.30 SD and 0.33 SD, respectively, although the estimate for Asians is not statistically significant.

We further explore the trust and rapport between teacher-evaluator pairs to better understand our findings on racial congruence. In Table 4 column 2, we add three measures that capture teachers' perceptions of mutual respect between themselves and their evaluator, their trust in their evaluator, and how much they enjoyed working with their evaluator.[7] We find that all three measures are positively associated with perceived evaluation feedback quality. For example, a one-point increase in a teacher's response for the trust measure (on the five-point Likert scale) is associated with over a quarter of a SD increase in perceived evaluation feedback quality. Notably, adding these variables into our model attenuates the point estimates for our teacher-evaluator race congruence dummy variables. This suggests that some of the positive relationship between race congruence and perceived evaluation quality can be explained by the higher likelihood that teachers of color feel a sense of mutual respect, trust, and enjoyment working with evaluators of the same race.

We conduct a more formal mediation analysis to understand how racial congruence is related to perceptions of evaluation feedback quality through trust, rapport and enjoyment.[8] Simple descriptive statistics show that teachers with evaluators that share their race are more likely than other teachers to feel mutual respect with their evaluator (78% vs 73%), to trust their

[7] The independent teacher survey included the following three items: 1) to what degree their relationship with their evaluator was characterized by mutual respect, 2) how much they trusted their evaluator, and 3) how much they enjoyed working with their evaluator.

[8] We look at the attenuation in our racial congruence estimates from model (1) to (2) and follow Preacher & Hayes (2008) to compare indirect effects in multiple mediator models. The attenuation in the point estimates as result of adding controls for trust and rapport allows us to see the proportion of the relationship between racial congruence and perceived evaluation feedback quality that is mediated via trust, rapport, and enjoyment.

evaluator (67% vs 58%), and to enjoy working with their evaluator (67% vs 60%). We find that over half of the variation between racial congruence and perceived evaluation feedback quality is explained by the measures of respect, trust, and enjoyment. For example, for African-American teachers and evaluators about half of the association between racial congruence and perceived evaluation feedback quality is mediated by the three measures. For Asians, this number is over 70% and for Hispanics it is over 80%. These findings are suggestive of the important role that relationships play in supporting perceptions of feedback quality. At the same time, our ability to infer the specific pathways through which these relationships operate are limited by the simultaneous measurement of these mediators and perceived evaluation feedback quality.

We conduct a number of robustness checks to see if the descriptive patterns we report above remain consistent across different model specifications. It is possible that the relationships we observe may simply reflect teachers' satisfaction with their summative evaluation rating since most teachers filled out surveys after receiving their summative ratings. Although we control for these ratings across all specifications, we further examine this potential threat by restricting our sample to only teachers who received the exact same rating of *Proficient.* As shown in column 3, we find strikingly similar results to those from our full sample after removing variation in summative evaluation ratings. In column 4 of Table 4, we show that our estimates remain qualitatively similar when we restrict our comparisons to teachers and evaluators within the same school by including school fixed effects. This suggests that unobserved time-invariant school characteristics are not driving the associations we find. Finally, in column 5, we apply weights to our preferred model based on the school-level teacher response rates to our independent teacher survey to test the sensitivity of our results to survey response bias. Again, our findings

remain quite similar suggesting that our estimates are not substantially biased by differential survey response rates across schools.

BPS Administrator Training Intervention

We next assess the implementation and effects of the BPS evaluator training series aimed at strengthening evaluators' ability to provide effective feedback. Evidence suggests that the district implemented the program with relatively high fidelity in both the years we studied. As we show in Table 5, 60% of evaluators randomly assigned to attend the program in the first year attended at least one of the 3-5 training sessions, with 40% attending all sessions in their series. Attendance increased slightly in the second year with 71% attending at least one session and 52% attending the full sequence.

Evaluators rated the training they received quite favorably and felt it would help them improve their evaluation feedback. Figure 5 illustrates how evaluators felt more capable at providing high-quality evaluation feedback after completing the training. Using measures on a nine-point Likert scale, 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) 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). 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.

Despite high attendance rates and positive feedback from evaluators, we find no effects of assigning evaluators to attend the training program on the perceived quality of evaluation feedback. In Table 6, we present intent-to-treat estimates of being assigned to attend the training in the first year on a range of outcomes. Our estimate of the intent-to-treat effect of the program on perceived evaluation feedback in the first year is small and statistically insignificant (-0.03 SD). In our preferred model, we do find evidence that the training had a marginally significant effect on the time between observations and post-observation feedback meetings, reducing it by 1.33 days relative to the control group mean of 5.34 days. However, we fail to find any significant effects on a range of evaluation implementation measures such as the number of observations, the number and length of post-observation meetings, or teachers' overall evaluation ratings. We also find relatively precise zeros for effects on teacher retention and student achievement.

The intervention does appear to have had a small negative effect on teachers' perceptions of school leadership quality, self-efficacy for classroom management, and self-efficacy for instructional strategies, though these effects vary depending on our definition of treatment. Our primary results suggest that the training intervention had a marginally significant 0.17 SD decrease in teachers' perceptions of the quality of school leadership, a significant 0.07 SD decrease in teachers' perceptions of their self-efficacy for classroom management, and a marginally significant 0.06 SD decrease in teachers' perceptions of their self-efficacy for classroom instruction. Although it is possible these results reflect real decreases in the effectiveness of principals' efforts and teachers' classroom management and instructional practices, they might also be the result of more critical evaluation feedback from administrators.

More critical feedback may have caused teachers to rate their school leadership quality lower and to recalibrate assessments of their own skills with behavior management.

We report results from our two alternative treatment-control contrasts in Appendix Table A4. We find no differential treatment effects based on the timing an evaluator attended the training during the school year on almost all of our outcomes of interest. We do find that teachers who were evaluated by evaluators that were randomly assigned to the fall versus the spring training sessions were less likely to return to their school (6 percentage point decrease) in the following year. This finding might reflect lower retention due to evaluators providing more critical feedback, and possibly their efforts to encourage more voluntary exits among teachers.

Overall, we find very few medium-term effects of the training program suggesting 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 find no evidence that the evaluation training improved teachers' perceptions of evaluators' feedback or other evaluation-related outcomes the year after evaluators completed the training. We do find a negative and statistically significant impact on teachers' self-reported classroom management practices and instructional strategies that may be a product of teachers' recalibrating assessments of their own instruction based on more meaningful feedback. 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 selected impact findings are the result of multiple hypothesis testing.

Discussion

In this study, we examine the efforts of BPS to promote professional development through their evaluation system. We find that evaluators were successful at evaluating teachers regularly and fairly, but that they struggled to provide feedback that teachers judged as high-quality. We observe both systematic variation across evaluators in the perceived quality of feedback they provided and considerable variation within evaluators across teachers. We also find that a range of teacher and evaluator characteristics – namely an evaluator and teacher being of the same race – are associated with perceived evaluation feedback quality.

The experimental results from our study suggests that training alone may not be sufficient in helping evaluators improve their ability to identify and communicate high-quality evaluation feedback. The null effects of the training program on even proximal outcomes are surprising considering the context of the intervention. The training was designed by experienced district-based evaluators to give participants opportunities to be active learners and try new approaches, tailored to the local context, grounded in adult learning practices, and pilot tested. Moreover, evaluators liked the training, thought it was of high quality, and intended to use practices learned during the evaluation process. However, our null results are consistent with a prior randomized control trial of a principal coaching program that found no effects on principals' ability to support teacher instructional development (Goff, Edward Guthrie, Goldring, & Bickman, 2014).

BPS has remained committed to refining and improving the evaluator training series in the years after we studied the program. Changes have focused on cutting extraneous information that was not aligned with core course objectives, providing more practical tools and exemplars, and substantially increasing the amount of in-class peer feedback and out-of-class feedback from training facilitators. We do find some encouraging evidence teachers' perceptions of evaluation feedback quality were improving over time as shown in Figure 1. Teachers reported receiving

feedback that was on average 0.06 SD higher quality in the second year of the training program relative to the first year. The percentage of teachers who agreed that their instruction improved because of the evaluation feedback they received increased from 24% to 27%. Our evaluation of the training program in its first full year of implementation fails to capture these positive overall trends in the implementation of the evaluation system.

At the same time, the training program was unable to address two cores constraints on evaluators' ability to provide in-depth high-quality feedback: evaluators' lack of time and limited content expertise. Capacity constraints limited evaluators' ability to find time to meet with teachers regularly. During training sessions many evaluators were candid about the fact that they would not have enough time to implement new evaluation feedback strategies with all of the teachers they were responsible for. Evaluators were, on average, conducting 50-60 classroom observations a year, with some doing up to four times as many. Conducting this large number of observations likely contributed to fewer post-observation meetings with teachers, shorter meetings, and longer wait times between observations and post-observation meetings – all of which are associated with lower-quality evaluation feedback (Kraft & Gilmour, 2016; Donaldson & Woulfin, 2018).

Time constraints may have also caused evaluators to prioritize elements of the evaluation system that district administrators could most closely monitor such as evaluation ratings and written feedback. Evaluators reported spending almost two-thirds of their evaluation time observing and analyzing data and writing evaluations and much less time having in-person discussions with teachers about how to improve their instruction. An incentive structure that does not track or reward in-person feedback is unlikely to result in frequent feedback conversations between evaluators and teachers.

The evaluator training program was also unable 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). Evaluators' lack of experience in multiple grades and subjects likely contributed to the large variability in how teachers perceived feedback quality and is not something that can easily be addressed through additional training.

Conclusion

The passage of ESSA has provided states and districts with broad flexibility in how they evaluate teachers. States and districts looking to revise teacher evaluation practices should carefully consider the alignment between their stated goals, system design, and resource investments. The theory of action underlying most teacher evaluation reforms posits that evaluation can improve student outcomes in two ways: 1) by differentiating among teacher performance to remove ineffectual teachers and attract and reward effective ones, and 2) by developing teachers and improving their instructional practices. States and districts that are committed to promoting teacher development through the evaluation process should consider closely the skills and capacity of those tasked with evaluating teachers as well as the supports available for them.

Consistent with prior research across several states, we find that teachers do not report regularly receiving high-quality feedback as part of the evaluation process. At the same time, the large variability we find in perceived evaluation quality across evaluators suggests that there are likely evaluators in every district that are successful at providing high-quality feedback. Districts

should seek to identify these evaluators to better understand their approaches to generating and communicating high-quality feedback. Likewise, districts might consider better leveraging effective evaluators through full-time instructional leadership or coaching roles rather than relying on administrators to drive instructional improvement. 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. Moreover, such a policy would free up time for principals to focus on fostering supportive environments where teachers feel comfortable and committed to ongoing professional improvement.

Districts might also focus on strengthening teacher-evaluator relationships and school cultures. There are no easy policy solutions for this work. Our findings suggest that diversity training might pay an important role in supporting constructive feedback conversations between teachers and evaluators of different races. The positive association between teacher-evaluator racial congruence and perceived evaluation feedback quality among teachers and evaluators of color also points to the importance of developing a diverse corps of evaluators.

Promoting instructional improvement through observation and feedback cycles is likely to be most successful when evaluators are instructional experts that develop strong relationships with teachers, when evaluators have the time to work intensively with teachers to provide in-depth feedback, when teachers perceive this feedback as high quality, and when teachers work in school environments where they are comfortable recognizing their weaknesses and committed 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 develop through the evaluation process.

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.

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

Campbell, S. L., & Ronfeldt, M. (2018). Observational Evaluation of Teachers: Measuring More Than We Bargained for?. *American Educational Research Journal*, 0002831218776216.

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*.

Cullen, J. B., Koedel, C., & Parsons, E. (2016). *The compositional effect of rigorous teacher evaluation on workforce quality* (No. w22805). National Bureau of Economic Research.

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.

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.

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

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., Woulfin, S., LeChasseur, K., & Cobb, C. D. (2016). The structure and substance of teachers' opportunities to learn about teacher evaluation reform: Promise or pitfall for equity?. *Equity & Excellence in Education*, *49*(2), 183-201.

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.

Egalite, A. J., & Kisida, B. (2018). The effects of teacher match on students' academic perceptions and attitudes. *Educational Evaluation and Policy Analysis*, *40*(1), 59-81.

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., 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.

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.

Goldring, E., Grissom, J. A., Rubin, M., Neumerski, C. M., Cannata, M., Drake, T., & Schuermann, P. (2015). Make room value added: Principals' human capital decisions and the emergence of teacher observation data. *Educational Researcher*, *44*(2), 96-104.

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.

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

Institute for Education Science. (2014). *State requirements for teacher evaluation policies promoted by Race to the Top* (NCEE Evaluation Brief). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

Jackson, C., & Cowan, J. (2018). *Assessing the Evidence on Teacher Evaluation Reforms* (CALDER Policy Brief No. 13-1218-1). American Institutes for Research/CALDER.

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

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.

Kraft, M.A., Brunner, E.J., Dougherty, S.M., & Schwegman, D. (2019) The effect of teacher evaluation reforms on new teacher supply and quality. Brown University Working Paper.

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

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. (2016). *Learning job skills from colleagues at work: Evidence from a field experiment using teacher performance data* (No. w21986). National Bureau of Economic Research.

Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior research methods*, *40*(3), 879 891.

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.

Ronfeldt, M., Farmer, S. O., McQueen, K., & Grissom, J. A. (2015). Teacher collaboration in instructional teams and student achievement. *American Educational Research Journal*, *52*(3), 475-514.

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

Sartain, L., & Steinberg, M. P. (2016). Teachers' labor market responses to performance evaluation reform: Experimental evidence from Chicago public schools. *Journal of Human Resources*, *51*(3), 615-655.

Chicago, IL: Consortium on Chicago School Research.

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., & Garrett, R. (2016). Classroom composition and measured teacher performance: What do teacher observation scores really measure?. *Educational Evaluation and Policy Analysis*, *38*(2), 293-317.

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.

Whitehurst, G., Chingos, M. M., & Lindquist, K. M. (2014). Evaluating teachers with classroom observations. *Brown Center on Education Policy: Brookings Institute*.

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

	Preferred Model	Adding Trust, Rapport, and Enjoyment Controls	Only teachers rated *Proficient*	School fixed effects	Weighted by teacher survey response rate
Teacher characteristics					
Age	0.011***	0.005***	0.010***	0.011***	0.011***
	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)
Female	-0.147***	-0.078***	-0.157***	-0.151***	-0.148***
	(0.044)	(0.029)	(0.047)	(0.045)	(0.043)
Graduate degree	-0.087**	-0.056**	-0.088**	-0.072*	-0.084**
	(0.038)	(0.024)	(0.042)	(0.036)	(0.039)
Experience					
3-5	-0.136**	-0.092*	-0.136*	-0.127*	-0.125*
	(0.066)	(0.054)	(0.070)	(0.064)	(0.064)
6-8	-0.274***	-0.110**	-0.253***	-0.269***	-0.292***
	(0.074)	(0.052)	(0.079)	(0.067)	(0.075)
9+	-0.328***	-0.104*	-0.311***	-0.335***	-0.350***
	(0.073)	(0.054)	(0.084)	(0.069)	(0.074)
Race/ethnicity					
African-American	0.192**	0.156***	0.244***	0.173**	0.211***
	(0.077)	(0.052)	(0.091)	(0.076)	(0.077)
Asian	0.249***	0.156**	0.308***	0.234***	0.250***
	(0.086)	(0.063)	(0.093)	(0.087)	(0.091)
Hispanic	0.131*	0.067	0.205**	0.148**	0.140*
	(0.075)	(0.050)	(0.083)	(0.073)	(0.076)
Summative rating					
Needs improvement	0.371**	0.114		0.291*	0.363**
	(0.159)	(0.107)		(0.162)	(0.177)
Proficient	1.392***	0.169		1.278***	1.401***
	(0.160)	(0.107)		(0.165)	(0.174)
Exemplary	1.639***	0.147		1.533***	1.647***
	(0.167)	(0.118)		(0.173)	(0.181)
Evaluator characteristics					
Age	-0.008***	-0.006***	-0.008**	-0.008**	-0.007**
	(0.003)	(0.002)	(0.003)	(0.004)	(0.003)
Female	0.163***	0.133***	0.157***	0.156***	0.168***
	(0.045)	(0.031)	(0.058)	(0.055)	(0.049)
Tenure at school					
3-5	-0.023	0.033	-0.049	0.073	-0.044
	(0.053)	(0.037)	(0.062)	(0.057)	(0.056)
6-8	0.158***	0.062	0.107	0.238***	0.147**
	(0.058)	(0.044)	(0.072)	(0.068)	(0.060)
9+	0.149	0.018	0.189	0.228**	0.113
	(0.100)	(0.090)	(0.121)	(0.102)	(0.110)
Race/ethnicity					
African-American	-0.219***	-0.153***	-0.184*	-0.186*	-0.225**
	(0.083)	(0.056)	(0.093)	(0.102)	(0.090)
Asian	-0.282*	-0.174*	-0.393**	-0.418**	-0.310*
	(0.157)	(0.088)	(0.169)	(0.165)	(0.177)
Hispanic	-0.094	-0.032	-0.052	-0.179**	-0.074

	(0.078)	(0.068)	(0.098)	(0.087)	(0.082)
Evaluator and teacher congruence					
Both same gender	0.010	0.028	-0.013	0.020	0.018
	(0.042)	(0.030)	(0.047)	(0.043)	(0.043)
Both African-American	0.293***	0.151**	0.289**	0.224**	0.267***
	(0.101)	(0.061)	(0.119)	(0.102)	(0.102)
Both Asian	0.330	0.112	0.572*	0.258	0.336
	(0.217)	(0.084)	(0.304)	(0.215)	(0.238)
Both Hispanic	0.304***	0.058	0.272**	0.344***	0.282**
	(0.111)	(0.067)	(0.133)	(0.122)	(0.109)
Both white	0.125	0.029	0.202**	0.100	0.130
	(0.078)	(0.057)	(0.093)	(0.076)	(0.083)
Respect[a]		0.101***			
		(0.019)			
Trust[a]		0.256***			
		(0.023)			
Enjoyment[a]		0.258***			
		(0.022)			
Survey response weights	N	N	N	N	Y
School fixed effects	N	N	N	Y	N
n	4,213	4,213	3,181	4,213	4,213

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 A3). 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]In the independent teacher survey, we asked teachers three questions: 1) to what degree their relationship with their evaluator was characterized by mutual respect, 2) how much they trusted their evaluator, and 3) how much they enjoyed working with their evaluator. Answers were on a five-point Likert scale that ranged from *Not at all* to *Completely* or A *tremendous amount.*

Table 5. *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 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 6. *The Effect of Evaluator Training on Teacher and Student Outcomes, Year 1 vs Year 2*

Outcomes	n	Uncontrolled	Controlled
Feedback Quality	2,033	-0.07	-0.03
		(0.08)	(0.06)
Ever met to discuss feedback	2,151	0.02	0.02
		(0.04)	(0.03)
Number of observations	2,094	0.17	0.26
		(0.38)	(0.38)
Number of discussion meetings	2,151	0.11	0.10
		(0.31)	(0.34)
Meeting length (minutes)	2,151	0.19	-0.39
		(1.13)	(1.05)
Time between observation and meeting (days)[a]	1,265	-0.61	-1.33*
		(0.66)	(0.75)
School leadership quality[b]	2,907	-0.31***	-0.17*
		(0.11)	(0.10)
Self-efficacy for classroom management[b]	2,907	-0.13***	-0.07**
		(0.05)	(0.03)
Self-efficacy for instructional strategies[b]	2,907	-0.06	-0.06*
		(0.04)	(0.04)
Summative Rating: Overall	3,904	0.03	0.02
		(0.04)	(0.03)
Summative Rating: Curriculum, Planning, and Assessment	3,904	0.02	0.03
		(0.04)	(0.03)
Summative Rating: Teaching All Students	3,904	0.02	0.03
		(0.04)	(0.03)
Summative Rating: Family and Community Engagement	3,904	-0.01	-0.01
		(0.03)	(0.02)
Summative Rating: Professional Culture	3,904	0.03	0.01
		(0.04)	(0.03)
Teacher retention for next year	3,904	-0.03	-0.03
		(0.03)	(0.03)
Student math achievement (no lagged outcome)[c]	53,664	-0.02	-0.00
		(0.14)	(0.08)
Student math achievement (lagged outcome)[c, d]	41,864	-0.04	-0.02
		(0.15)	(0.03)
Student ELA achievement (no lagged outcome)[c]	53,056	0.03	0.05
		(0.12)	(0.06)
Student ELA achievement (lagged outcome)[c, d]	41,355	0.05	0.04
		(0.12)	(0.03)

Notes: *** p<0.01, ** p<0.05, * p<0.1. Standard errors are in parenthesis and 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, 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 by Massachusetts Comprehensive Assessment System (MCAS) test scores. The MCAS is a statewide exam administered to students in grades 3-10 in mathematics and ELA. We standardize scores at the year, grade, subject level to have a mean of 0 and standard deviation of 1. We include one year lagged achievement outcomes as controls. Out of the 123 schools in our sample, test score data is available for only 115 schools. Test score data was unavailable for particular schools, such as those that were focused on children with disabilities or off-track students.

[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.

Figures

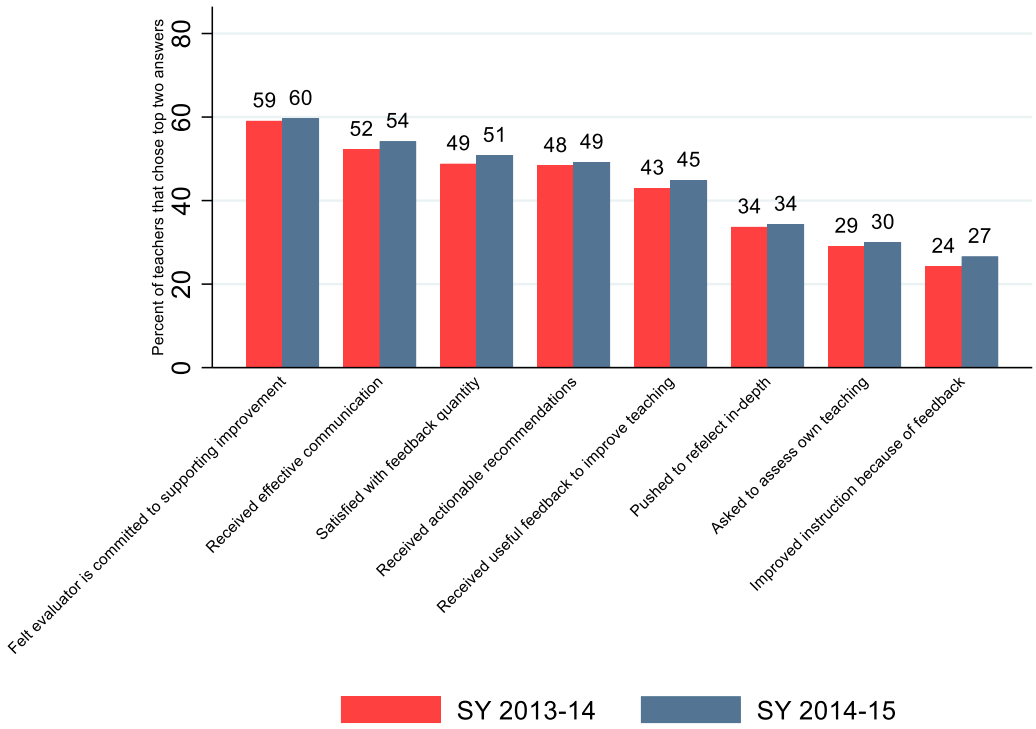

Figure 1. Agreement rates for items included in the perceived quality of evaluation feedback scale for the 2013-14 and 2014-15 school years

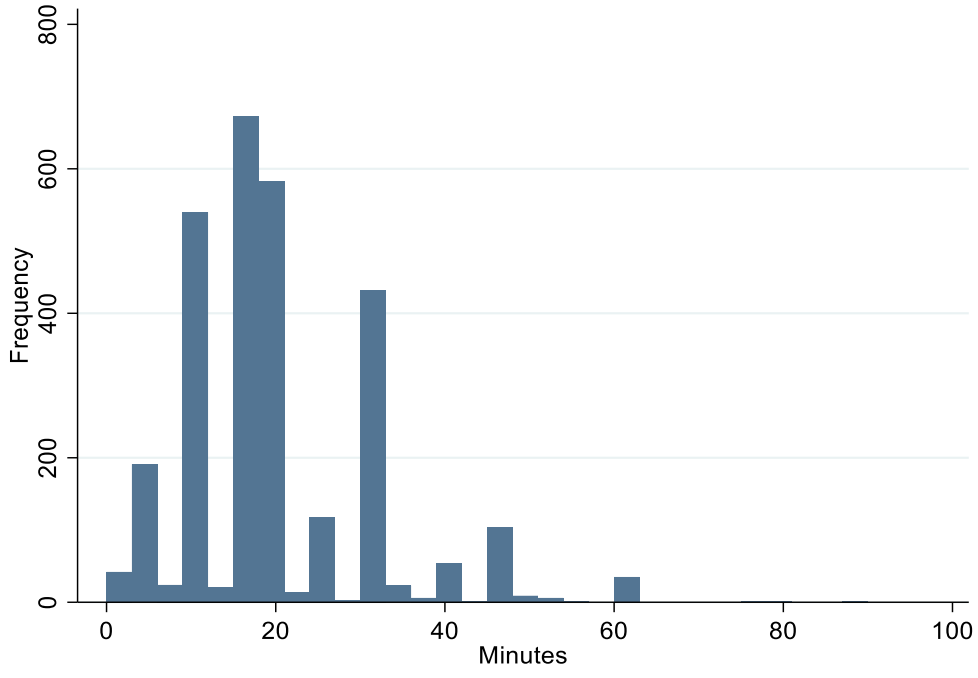

Figure 2. The length of post-observation meetings across the 2013-14 and 2014-15 school years.

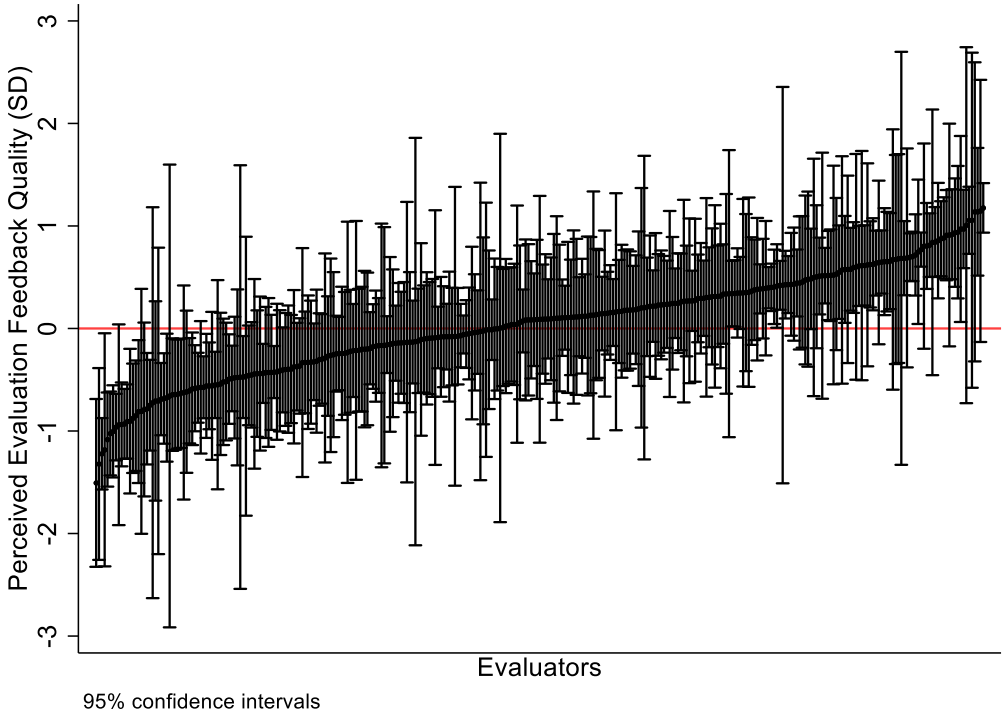

Figure 3. Distribution of average perceived evaluation feedback quality across evaluators for the 2013-14 and 2014-15 school years.
Notes: This figure is subset to evaluators who evaluated at least five teachers and only shows evaluators whose confidence intervals are between -3 SD and 3 SD. This excludes 23 evaluators.

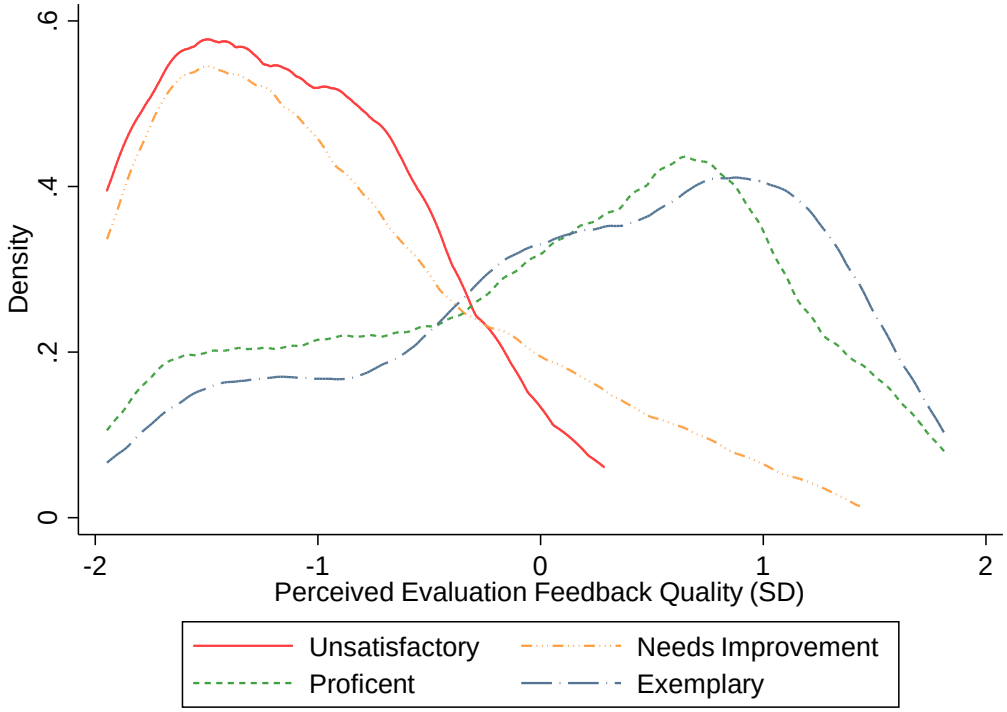

Figure 4. Distribution of perceived evaluation feedback quality by summative teacher evaluation ratings, from the 2013-14 and 2014-15 school years.

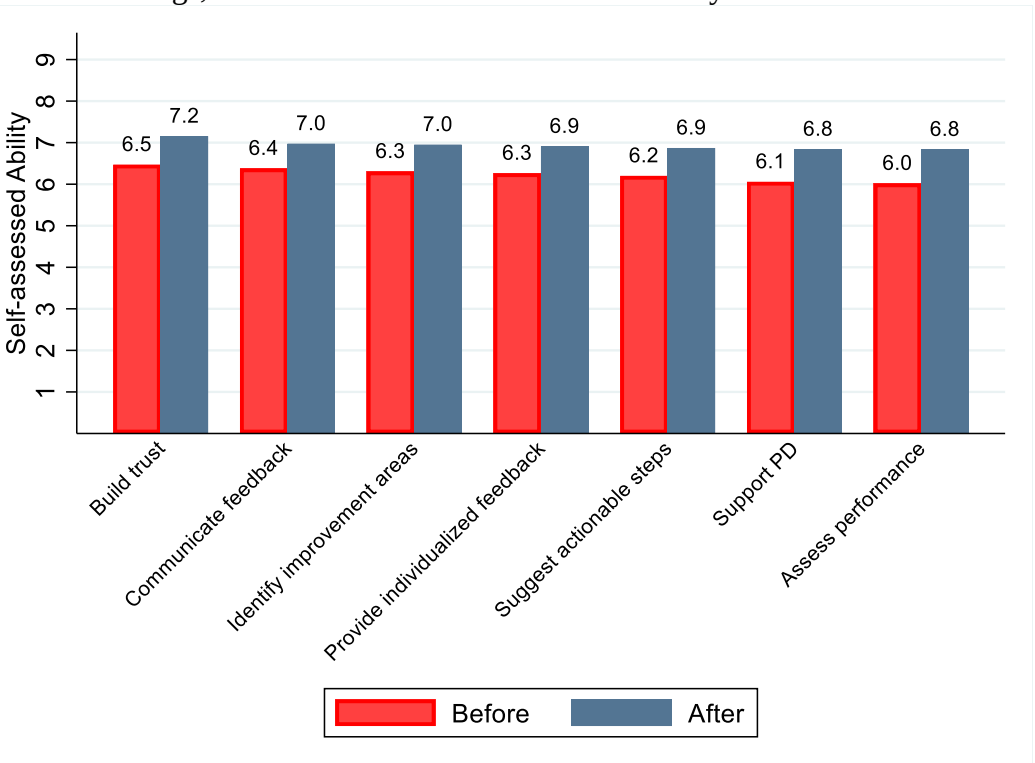

Figure 5. Evaluators' self-assessment of evaluation skills pre- and-post training across the 2013-14 and 2014-15 school years.

Appendix Tables

Table A1. *Items Included in the Perceived Evaluation Feedback Quality Scale*

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 A2. *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 school climate survey, which are measures of the school climate.

[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.

Table A3. *The Relationship Between School Characteristics and Perceived Evaluation Feedback Quality*

	Preferred Model	Adding Trust, Rapport, and Enjoyment Controls	Only teachers rated *Proficient*	School fixed effects	Weighted by teacher survey response rate
School characteristics					
Prior year school climate					
Collegial work environment	-0.042	-0.035	-0.068	-0.153**	-0.043
	(0.045)	(0.037)	(0.052)	(0.069)	(0.049)
School leadership quality	0.160***	0.026	0.170***	0.016	0.167***
	(0.032)	(0.020)	(0.038)	(0.052)	(0.035)
Parent and student engagement	-0.126*	-0.044	-0.139*	-0.158	-0.146**
	(0.069)	(0.040)	(0.078)	(0.130)	(0.072)
Collective teacher efficacy	-0.051	0.019	-0.027	0.084	-0.067
	(0.055)	(0.042)	(0.062)	(0.082)	(0.059)
Self-efficacy for classroom management	0.035	0.043	0.014	0.038	0.045
	(0.048)	(0.030)	(0.052)	(0.089)	(0.049)
Teacher influence over classroom decision-making	-0.062**	-0.043**	-0.070**	0.048	-0.065**
	(0.025)	(0.019)	(0.031)	(0.074)	(0.027)
Self-efficacy for instructional strategies	0.078**	0.033	0.059	0.009	0.087**
	(0.033)	(0.024)	(0.037)	(0.046)	(0.036)
Relationship with students and parents	0.073*	-0.002	0.085*	-0.036	0.079*
	(0.041)	(0.028)	(0.047)	(0.079)	(0.044)
Enrollment	-0.000	-0.000	-0.000	-0.001	-0.000
	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)
Student to teacher ratio	0.018	0.006	0.007	0.018	0.021
	(0.012)	(0.007)	(0.014)	(0.021)	(0.013)
Percent of high needs students	-0.000	0.000	-0.005	-0.011*	0.000
	(0.004)	(0.003)	(0.005)	(0.006)	(0.004)
Percent of ELL students	-0.002	-0.002	-0.003	-0.002	-0.002
	(0.002)	(0.002)	(0.003)	(0.010)	(0.003)
Percent of students with disabilities	0.004**	0.001	0.005*	-0.002	0.004*
	(0.002)	(0.001)	(0.002)	(0.008)	(0.002)
Percent of students African-American	-0.005**	-0.002	-0.002	-0.014**	-0.005*

	(0.002)	(0.002)	(0.003)	(0.007)	(0.003)
Percent of students Asian	-0.008**	0.001	-0.002	-0.005	-0.008*
	(0.004)	(0.003)	(0.005)	(0.024)	(0.004)
Percent of students Hispanic	-0.003	0.001	0.003	0.000	-0.003
	(0.003)	(0.002)	(0.004)	(0.011)	(0.004)
Percent of students Other	-0.003	0.003	0.018	0.025	0.001
	(0.023)	(0.016)	(0.026)	(0.052)	(0.026)
Survey response weights	N	N	N	N	Y
School fixed effects	N	N	N	Y	N
n	4,213	4,213	3,181	4,213	4,213

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 4 – for the estimates for teacher and evaluator characteristics refer to Table 4. All models contain fixed effects for school year. Standard errors are clustered at the school level.

Appendix A4. *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	2,103	-0.05	-0.07	1,667	-0.09	-0.09
		(0.08)	(0.07)		(0.09)	(0.08)
Ever met to discuss feedback	2,210	-0.01	-0.02	1,729	-0.03	-0.03
		(0.04)	(0.03)		(0.05)	(0.05)
Number of observations	2,183	0.34	0.11	1,721	-0.13	-0.28
		(0.37)	(0.39)		(0.34)	(0.37)
Number of discussion meetings	2,210	0.17	0.01	1,729	-0.21	-0.22
		(0.31)	(0.26)		(0.28)	(0.27)
Meeting length (minutes)	2,210	-1.26	-1.13	1,729	-1.37	-1.23
		(1.02)	(0.99)		(1.37)	(1.31)
Time between observation and meeting (days)[a]	1,369	0.10	0.55	1,110	0.13	-0.28
		(0.40)	(0.43)		(0.51)	(0.54)
School leadership quality[b]	3,178	-0.09	-0.16*	2,528	-0.15	-0.10
		(0.11)	(0.09)		(0.11)	(0.11)
Self-efficacy for classroom management[b]	3,178	0.02	-0.08**	2,528	-0.12*	-0.14**
		(0.05)	(0.04)		(0.07)	(0.06)
Self-efficacy for instructional strategies[b]	3,178	-0.05	-0.04	2,528	-0.21***	-0.15**
		(0.05)	(0.03)		(0.06)	(0.06)
Summative Rating: Overall	3,831	0.01	-0.02	2,840	0.02	0.01
		(0.04)	(0.03)		(0.03)	(0.02)
Summative Rating: Curriculum, Planning, and Assessment	3,831	0.01	-0.03	2,840	0.00	-0.01
		(0.04)	(0.03)		(0.04)	(0.03)
Summative Rating: Teaching All Students	3,831	0.01	-0.03	2,840	0.00	-0.01
		(0.04)	(0.03)		(0.04)	(0.03)
Summative Rating: Family and Community Engagement	3,831	-0.00	-0.03	2,840	0.01	0.02

		(0.03)	(0.02)		(0.03)	(0.03)
Summative Rating: Professional Culture	3,831	-0.02	-0.05	2,840	-0.00	0.00
		(0.04)	(0.03)		(0.03)	(0.03)
Teacher retention for next year	3,831	-0.04	-0.06**	2,840	-0.03	-0.03
		(0.03)	(0.03)		(0.04)	(0.04)
Student math achievement (no lagged outcome)[c]	51,230	-0.02	-0.04	36,779	-0.04	-0.04
		(0.12)	(0.07)		(0.16)	(0.09)
Student math achievement (lagged outcome)[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 (no lagged outcome)[c]	51,013	-0.04	-0.05	36,811	-0.09	-0.08
		(0.10)	(0.06)		(0.15)	(0.08)
Student ELA achievement (lagged outcome)[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 by Massachusetts Comprehensive Assessment System (MCAS) test scores. The MCAS is a statewide exam administered to students in grades 3-10 in mathematics and ELA. We standardize scores at the year, grade, subject level to have a mean of 0 and standard deviation of 1. Out of the 123 schools in our sample, test score data is available for only 115 schools. Test score data was unavailable for particular schools, such as those that were focused on children with disabilities or off-track students.

[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 21-23% of our sample.