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

Matthew A. Kraft & Alvin Christian

Background

Teacher evaluation reforms over the past decade have been motivated by differentiating teacher performance for accountability and promoting professional development through classroom observations and feedback. A growing body of evidence shows the potential of frequent feedback to improve instructional practices and student achievement, yet we know little of what drives high-quality feedback.

Purpose

In this study, we examine teachers' perceptions of feedback received through Boston Public Schools' (BPS) new teacher evaluation system and evaluate the district's efforts to improve feedback quality. We also explore what teacher, evaluator, and school characteristics are correlated with feedback that teachers perceived as high quality.

Setting

In the 2011-12 academic year, BPS reformed its evaluation system and convened a group of experienced administrators to develop and pilot a multi-day evaluator training program to improve the quality of feedback provided to teachers. We evaluate the implementation and effects of this training series by exploiting the staggered rollout of the program across two academic years.

Participants

Our sample of 123 BPS schools includes traditional public schools, charter schools, and pilot schools (see **Table 1)**. During the study period, BPS employed 4,805 teachers (see **Table 2**) and 355 evaluators - principals, vice principals, and other school leaders (see **Table 3**).

Intervention

Alongside BPS, we developed a training series for evaluators. Several features differentiate this training from traditional professional development courses: 1) the training was taught by BPS school leaders instead of central office staff or external consultants; 2) the course featured guiding philosophies and theories and practical strategies; 3) participants practiced with homework between sessions and received individualized feedback on assignments; 4) finally, the training was intensive and occurred in small groups, consisting of 3-5 sessions totaling 15 hours with a cohort of 20-30 peers.

Research Design

Resource limitations required BPS to stagger program implementation over two years. This allowed us to randomize schools to training sessions in a given semester. We grouped schools into six blocks based on size (small and large) and type (elementary, middle, and high) and randomized within school size-type blocks. Schools could then choose their evaluator to attend a training series offered at different times each semester. 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.

Data Collection and Analysis

We obtained administrative data from BPS and conducted an independent survey of teachers and evaluators to identify their perceptions of the observation and feedback system.

We explore the relationship between perceived 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:

$$Evaluation\ Feedback_{ist} = \alpha + \beta X_{ist} + \gamma_t + \varepsilon_{ist}$$

Here X represents a vector of teacher, evaluator, and school characteristics. We include fixed effects for year, γ, and cluster standard errors at the school level.

We estimate the effects of random assignment to the training program during the 2013-14 school year on a range of outcomes using the following OLS model:

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

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 treatment. We control for teacher/student, evaluator, and school characteristics in X and use fixed effects for year, γ, and school size-type blocks, π. We cluster standard errors at the school level.

Findings

Teachers generally thought that evaluators were fair and accurate, but had less favorable views about the quality of feedback they received. In **Figure 1**, we show that only half of teachers surveyed reported satisfaction with the quantity of feedback received, and less than half of teachers found the feedback useful or actionable.

A variance decomposition of perceived feedback quality across and within evaluators reveals considerable heterogeneity across evaluators in providing effective feedback, as seen in **Figure 2**.

We next explore the relationship between teacher, evaluator, and school characteristics and teachers' perceptions of feedback quality. We show that less experienced teachers report receiving higher-quality feedback than their more experienced peers and that being a teacher of

color is associated with 0.13-0.25 SD higher reported evaluation feedback quality relative to white teachers. Teachers rate evaluators with more experience as providing higher-quality feedback and that compared to white evaluators, the perceived quality of evaluation feedback is lower for evaluators of color. School characteristics are weakly correlated with 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 rate feedback quality 0.29 SD higher than racially incongruent pairs. We find similar estimates for racial congruence among Hispanics and Asians. We also find that over half of the variation between racial congruence and perceived evaluation feedback quality is explained by measures of respect, trust, and enjoyment.

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 feedback, teacher retention, or student achievement. The intervention had negative effect on teachers' perceptions of school leadership quality (-0.48 SD), self-efficacy for classroom management (-0.20 SD), and self-efficacy for instructional strategies (-0.07 SD).

Conclusions

Training alone may be insufficient to improve evaluators' ability to identify and communicate high-quality evaluation feedback. Improving instruction through observation and feedback is likely to be most successful when evaluators are instructional experts that develop strong relationships with teachers, when evaluators have 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 development through the evaluation process.

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.

Figures

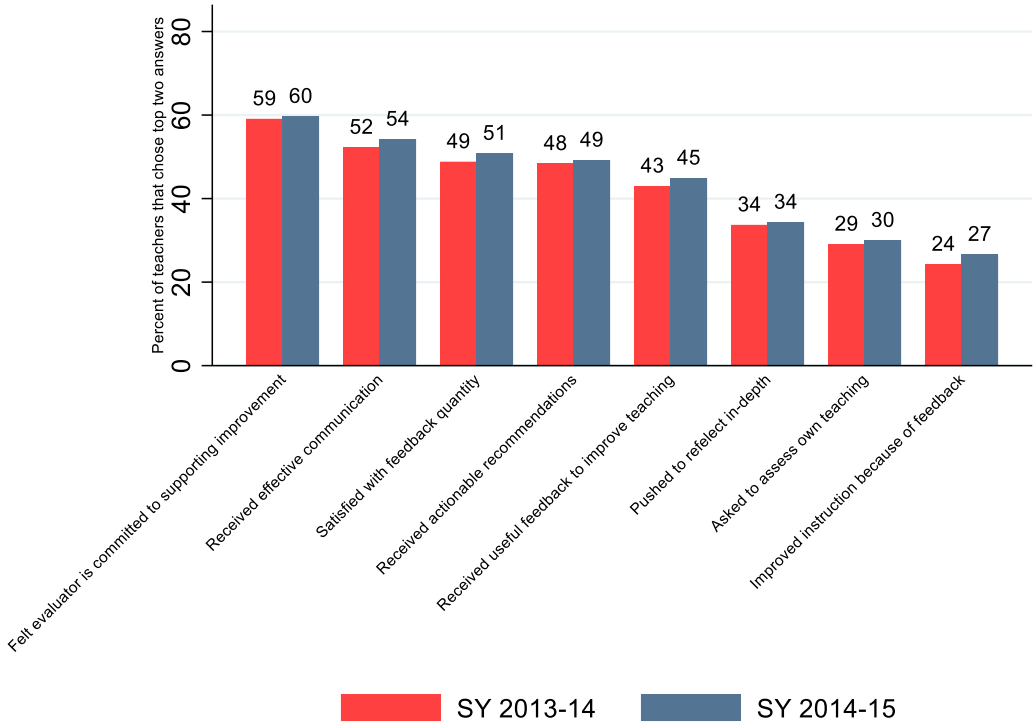

Figure 1. Perceived quality of evaluation feedback for the 2013-14 and 2014-15 school years.

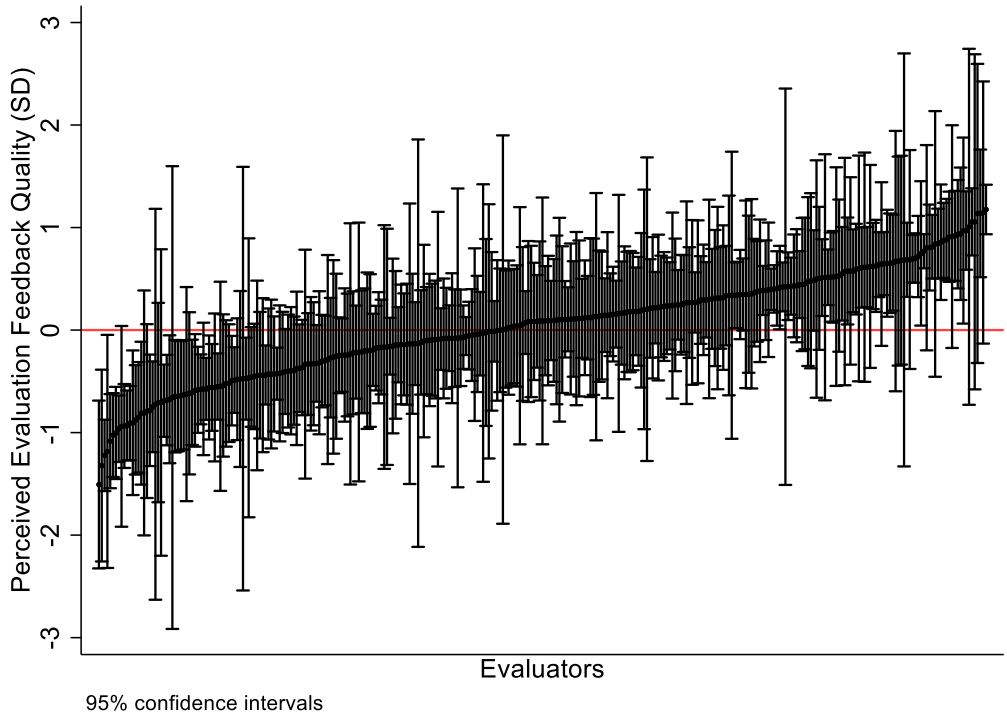

Figure 2. Distribution of perceived evaluation feedback quality at evaluator level across the 2013-14 and 2014-15 school years.
Notes: This figure is subset to evaluators who evaluated at least five teachers and only shows evaluators whose confidence intervals are between -3 SD and 3 SD. This excludes 23 evaluators.