U.S. Department of Education
December 2017

The Impact of Providing Performance Feedback to Teachers and Principals

Executive Summary

Michael S. Garet
Andrew J. Wayne
Seth Brown
Jordan Rickles
Mengli Song
David Manzeske
American Institutes for Research

Melanie Ali
Project Officer
Institute of Education Sciences

ies INSTITUTE OF EDUCATION SCIENCES

This page has been left blank for double-sided copying

The Impact of Providing Performance Feedback to Teachers and Principals

Executive Summary

December 2017

Michael S. Garet
Andrew J. Wayne
Seth Brown
Jordan Rickles
Mengli Song
David Manzeske
American Institutes for Research

Melanie Ali
Project Officer
Institute of Education Sciences

NCEE 2018-4000
U.S. DEPARTMENT OF EDUCATION

NATIONAL CENTER FOR EDUCATION EVALUATION AND REGIONAL ASSISTANCE
Institute of Education Sciences

This page has been left blank for double-sided copying.

U.S. Department of Education
Betsy DeVos
Secretary

Institute of Education Sciences
Thomas W. Brock
Commissioner, National Center for Education Research
Delegated Duties of the Director

National Center for Education Evaluation and Regional Assistance
Ricky Takai
Acting Commissioner

December 2017

This report was prepared for the Institute of Education Sciences under Contract ED-IES-11-C-0066. The project officer is Melanie Ali in the National Center for Education Evaluation and Regional Assistance.

IES evaluation reports present objective information on the conditions of implementation and impacts of the programs being evaluated. IES evaluation reports do not include conclusions or recommendations or views with regard to actions policymakers or practitioners should take in light of the findings in the report.

While permission to reprint this publication is not necessary, the citation should be:

Garet, M.S., Wayne, A.J., Brown, S., Rickles, J., Song, M., and Manzeske, D. (2017). *The Impact of Providing Performance Feedback to Teachers and Principals*, *Executive Summary* (NCEE 2018-4000). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

This report is available on the IES website at http://ies.ed.gov/ncee.

Alternate Formats: Upon request, this report is available in alternate formats, such as Braille, large print, audiotape, or computer diskette. For more information, please contact the Department’s Alternate Format Center at 202-260-9895 or 202-205-8113.

Acknowledgments

This study was a collaborative effort and involved a diverse group of partners. We were fortunate to have had the advice of our expert technical working group. Members included Thomas Cook, Northwestern University; Thomas Dee, Stanford University; Laura Goe, Educational Testing Service; Laura Hamilton, RAND; Daniel McCaffrey, Educational Testing Service; Catherine McClellan, Clowder Consulting; Jonah Rockoff, Columbia University; Carla Stevens, Houston Independent School District; John Tyler, Brown University; and Judy Wurtzel, Charles and Lynn Schusterman Foundation.

We would also like to thank all those who provided the teacher and principal performance feedback systems and training, including the organizations that supported the implementation of the Classroom Assessment Scoring System (the University of Virginia and Teachstone), the Framework for Teaching (the Danielson Group and Teachscape), and the VAL-ED (Discovery Education). We appreciate the willingness and commitment of the school district leaders, schools, principals, study-hired observers, and teachers to implement the intervention and data collection activities, which involved a significant amount of time and energy.

We are also grateful to the AIR staff who worked diligently to coordinate the study's performance feedback activities in participating districts: Rebecca Herman, Fran Stancavage, Matthew Clifford, Mariann Lemke, Susan Ward, Carmen Martinez, Muna Shami, Ben Kalina, Marlene Darwin, Carla Hulce, Nicol Christie, Debbie Davidson-Gibbs, Mark Garabaldi, Jessica Milton, Elaine Liebesman, Amy Potemski, Roshni Menon, Marian Eaton, Gur Hoshen, Zhongjie Sun, and Michele Cadigan. Additional AIR staff worked tirelessly on data collection: Dorothy Seidel, Lauren Staley, Cheryl Pruce, Sarah Bardack, Makeda Amelga, and Lindsey Mitchell. For their efforts to identify the partner districts, we thank the recruitment leaders Anja Kurki and Rebecca Herman and the team of senior recruiters: Kirk Walters, James Taylor, Marlene Darwin, Nicholas Sorensen, Mark Garibaldi, Carmen Martinez, Nicol Christie, Kathleen Perez-Lopez, and Emily Rosenthal. The study authors are also grateful to Rachel Garrett, Jenifer Harr-Robins, Luke Keele, and Paul Bailey for their help with data analyses, in addition to Connie Conroy who provided administrative assistance throughout the project. We also thank Jeanette Moses for her meticulous work supporting the development of the study reports, making sure they were ready for each stage of review and publication.

Finally, numerous staff from Instructional Research Group (IRG) and AIR coded the classroom recordings. We are especially grateful to Russell Gersten, Joe Dimino, and Mary Jo Taylor, who led IRG's coding efforts.

This page has been left blank for double-sided copying.

Disclosure of Potential Conflicts of Interest

The research team was comprised of staff from American Institutes for Research (AIR). None of the research team members has financial interests that could be affected by findings from The Impact of Providing Performance Feedback to Teachers and Principals. No one on the 10-member technical working group, convened by the research team three times to provide advice and guidance, has financial interests that could be affected by findings from the evaluation.

This page has been left blank for double-sided copying.

Executive Summary

Educator performance evaluation systems are a potential tool for improving student achievement by increasing the effectiveness of the educator workforce.[1] For example, recent research suggests that giving more frequent, specific feedback on classroom practice may lead to improvements in teacher performance and student achievement.[2]

This report is based on a study that the U.S. Department of Education's Institute of Education Sciences conducted on the implementation of teacher and principal performance measures that are highlighted by recent research, as well as the impact of providing feedback based on these measures.[3] As part of the study, eight districts were provided resources and support to implement the following three performance measures in a selected sample of schools in 2012–13 and 2013–14:

- *Classroom practice measure:* A measure of teacher classroom practice with subsequent feedback sessions conducted four times per year based on a classroom observation rubric.
- *Student growth measure:* A measure of teacher contributions to student achievement growth (i.e., value-added scores) provided to teachers and their principals once per year.
- *Principal leadership measure:* A measure of principal leadership with subsequent feedback sessions conducted twice per year.

Within each district, schools were randomly assigned to implement the performance measures (the treatment group) or not (the control group). No formal "stakes" were attached to the measures—for example, they were not used by the study districts for staffing decisions such as tenure or continued employment.[4] Instead, the measures were used to provide educators and their supervisors with information regarding performance. Such information might identify educators who need support and indicate areas for improvement, leading to improved classroom practice and leadership and boosting student achievement.

This is the second of two reports on the study. The first focused on the first year of implementation, describing the characteristics of the educator performance measures and teachers' and principals' experiences with feedback.[5] This report examines the impact of the two-year intervention, as well as implementation in both years. The main findings are:

- **The study's measures were generally implemented as planned.** For instance, teachers in treatment schools received an average of 3.7 and 3.9 observations with feedback sessions in Years 1 and 2, respectively. Almost all (98 percent) treatment teachers with

[1] See Stecher et al. (2016); Weisburg et al. (2009).

[2] See Steinberg and Sartain (2015); Taylor and Tyler (2012).

[3] For recent research on performance measures, see, for example, Bill & Melinda Gates Foundation (2012, 2013).

[4] There were exceptions in three districts. In these districts, the observations conducted by principals as part of the study counted in their official rating system if the teacher was due to be observed that year under the district's existing evaluation system.

[5] See Wayne et al. (2016).

value-added scores received printed student growth reports in Year 2, although less than half (39 percent) accessed their reports in Year 1, when disseminated online only.

- **The study's measures provided some information to identify educators who needed support, but provided limited information to indicate the areas of practice educators most needed to improve.** For example, although a large majority of teachers (more than 85 percent) had overall classroom observation scores in the top two performance levels, scores averaged over the year provided some reliable information to distinguish teacher performance (with Year 2 reliabilities of .53 to .61 and .70 to .77 for the two observation rubrics used). Differences in teachers' observation ratings across dimensions, however, had limited reliability to identify areas for improvement, even when averaged over the year (with Year 2 reliabilities of .35 to .43 and .18 to .30 for the two observation rubrics). Observation score reliabilities were similar in Year 1.
- **As intended, teachers and principals in treatment schools received more frequent feedback with ratings than teachers and principals in control schools.** Treatment teachers reported receiving more feedback sessions on their classroom practice with ratings and a written narrative justification than control teachers (3.0 versus 0.7 sessions, based on responses to a teacher survey in the spring of Year 1, and 3.0 versus 0.2 sessions in Year 2). Treatment principals received more instances of oral feedback with ratings on their leadership than control principals (1.0 versus 0.4 sessions based on responses to a principal survey in the spring of Year 1, and 2.0 versus 1.0 sessions reported at the end of Year 2).
- **The intervention had some positive impacts on teachers' classroom practice, principal leadership, and student achievement.** To assess the impact on classroom practice, the study team video-recorded lessons in both treatment and control schools and coded them with the two observation rubrics used to provide feedback. The intervention had a positive impact on teachers' classroom practice on one of the two observation rubrics, moving teachers from the 50th to the 57th percentile, but it had no impact on practice as measured by the other rubric. The intervention also had a positive impact on the two measures of principal leadership examined—instructional leadership and teacher-principal trust—moving teachers from the 50th to the 60th percentile on teacher-principal trust in Year 1, for example. In Year 1, the intervention had a positive impact on students' achievement in mathematics, amounting to about four weeks of learning. In Year 2, the impact on mathematics achievement was similar in magnitude but not statistically significant. The intervention did not have a statistically significant impact on reading/English language arts achievement in either year.

Study Overview

The study addressed five research questions:

1. To what extent were the performance measures and feedback implemented as planned?
2. To what extent did the performance measures identify more and less effective educators and signal dimensions of practice that most needed improvement?
3. To what extent did educators' experiences with performance feedback differ for treatment and control schools?

4. Did the intervention have an impact on teacher classroom practice and principal leadership?
5. Did the intervention have an impact on student achievement?

Study Design

The study used an experimental design in eight purposefully selected districts. We recruited districts that met the following criteria: (1) had at least 20 elementary and middle schools, (2) had data systems that were sufficient to support value-added analysis, and (3) had current performance measures and feedback that were less intensive than that implemented as part of the study. The recruited districts required fewer than four observations of teachers per year and did not require the inclusion of student achievement information in teacher ratings as part of their evaluation systems. None of the recruited districts used a principal leadership measure similar to that used by the study.

The study used two different classroom observation measures to provide feedback, to make the findings more broadly relevant than they would be if only one measure was used. Four of the eight districts used the Classroom Assessment and Scoring System (CLASS) and the other four used Charlotte Danielson's Framework for Teaching (FFT). The observation rubrics were not randomly assigned; districts chose based on preference. Thus, differences in the results in the CLASS and FFT districts cannot necessarily be attributed to the observation systems; differences could occur due to other district characteristics.

Each study district identified a set of regular elementary and middle schools willing to participate. In these schools, the study focused on the teachers of reading/English language arts and mathematics in grades 4–8, as well as the principals.[6] Both the treatment and the control schools continued to implement their district's existing performance evaluations and measures, and the treatment schools additionally implemented the study's performance measures with feedback. In total, 63 treatment schools and 64 control schools participated in the study.

Consistent with the recruitment criteria, the study districts were larger and more likely to be urban than the average U.S. district. The study schools were similar to schools in the national population in terms of enrollment and Title I status, but on average had a higher percentage of students who were minorities.

Data Sources

The study collected the following data on the performance feedback provided to teachers and principals in the treatment schools:

Implementation of the measures. We documented attendance at orientation and training events related to the study's performance measures. We also gathered data from the online systems maintained by the vendors on the frequency of classroom observations and feedback

[6] Teachers of kindergarten through grade 3 also participated in the study. This was done mainly to promote schoolwide engagement in the implementation of the classroom practice and principal leadership performance measures. These teachers were not included in the main study analyses, however, because student assessment data were not available for kindergarten through grade 3.

sessions, and teachers' and principals' access of student growth reports. Finally, surveys of teachers and principals administered in the spring of Year 2 included items for treatment group members that asked about their perceptions of the intervention. Principals and teachers in treatment schools reported on their perceptions of the performance information they received from the study's classroom observation and principal leadership practices measures compared to that received from the districts' official performance system.

Information provided to teachers and principals. We also collected the ratings generated by the teacher classroom practice, student growth, and principal leadership performance measures.

In addition, data were collected on the following teacher and principal experiences and initial outcomes in both treatment and control schools:

- **Educators' experiences with performance feedback.** In the spring of each study year, we surveyed the teachers and principals in treatment and control schools to collect information on the performance information educators received.
- **Educators' interest in improving.** The spring surveys also asked about initial outcomes, including whether teachers and principals wished to improve or sought professional development in areas covered by the feedback.

Finally, we collected data on three types of main outcomes in treatment and control schools:

- **Teachers' classroom practice.** In the spring of Year 2, to provide a common outcome measure, we video-recorded one lesson per teacher and then selected a random sample of half of the respondents for a second round of recording.[7] We coded each of the videos using the CLASS and FFT.[8] This allowed us to examine impacts on a measure of practice aligned with the measure used for feedback in the district's treatment group and a measure that was similar, but not completely aligned with that used for feedback in the district.
- **Principal leadership.** We relied on teacher responses on survey items designed to capture principal instructional leadership and teacher-principal trust, based on scales developed by the Chicago Consortium on School Research (CCSR 2012).
- **Student achievement.** We collected students' scores on state standardized tests in reading/English language arts and mathematics in each study year.

In addition to the information described above, we collected data on the characteristics of principals, teachers, and students in study schools from district administrative records.

[7] We video recorded two lessons for some teachers and one for others to achieve the desired precision while minimizing cost and burden.

[8] To the extent possible, video-recording was scheduled to take place when a teacher was teaching either reading/English language arts or mathematics. Overall, 45 percent of the video-recorded lessons were in reading/English language arts, 50 percent in mathematics, and 5 percent in other subjects.

Analyses

To examine the implementation of the teacher and principal performance measures, we analyzed the extent to which participants received the intended training on the measures, carried out the anticipated performance measurement activities, and received performance information and feedback as planned. We also examined the ratings teachers and principals received, including whether the ratings distinguished between lower and higher performers.

To assess whether the intervention led to differences between treatment and control schools in educators' experiences with performance measurement and feedback, and whether it led to changes in educator practice, we compared responses of teachers and principals in the treatment and control schools on the survey and ratings of teachers' practice based on video-recordings of their instruction. We also compared student achievement in reading/English language arts and mathematics in treatment and control schools. Finally, to supplement the impact analyses, we examined the association of classroom practice and principal leadership with student achievement.

Implementation of the Intervention

The intervention provided teachers and principals with information based on three performance measures: the first focused on teacher classroom practice, the second on student growth, and the third on principal leadership. The intervention was intended to provide teachers and principals frequent, systematic feedback to identify educators who need support and to signal specific areas of practice for improvement.

How well was the classroom practice measure implemented and what information did the measure provide?

The classroom practice component was designed to provide information on multiple dimensions of practice, based on observations conducted during four "windows" each year. One observation a year was to be conducted by a school administrator and three by observers hired by the study.[9] After each observation, the observer was to prepare a standard report with both ratings and narrative justification and to discuss the report with the teacher during a feedback session. The CLASS reports described classroom practice on 12 dimensions. Each dimension was scored on a 7-point scale and assigned a performance level (*ineffective*, *developing effectiveness*, *effective*, or *highly effective*). The CLASS also provided an overall score. The FFT described practice on up to 10 dimensions. Each dimension was scored on a 4-point scale (*unsatisfactory*, *basic*, *proficient*, or *distinguished*).

On average, teachers received nearly the four intended feedback sessions each year. The average number of feedback sessions per teacher was 3.7 in Year 1 and 3.9 in Year 2.

[9] To the extent possible given the constraints of scheduling, the principal and study-hired observers were asked to conduct the four observations for each teacher when the teacher was teaching the same subject (either reading/English language arts or mathematics) and during the same class period. Conducting observations during the same subject and class period was intended to make it easier for teachers and principals to interpret the observation ratings. In addition, within each school, the study-hired observers were encouraged to balance the number of teachers who were observed during reading/English language arts and mathematics, if feasible.

Teachers present in the spring of Year 2 received an average of 6.8 feedback sessions across the two years, instead of the intended eight sessions, primarily due to teacher mobility.

Nearly all teachers had classroom observation overall scores in the top two performance levels, limiting the potential of the information to signal a need for teachers to improve. For CLASS, in Year 2, for example, 98 percent or more of the teacher ratings within an observation window were in the top two of the four CLASS performance levels. For FFT, more than 87 percent of the teachers within an observation window had an overall score of 2.50 or higher, which corresponds to the top two of four study-defined performance levels.[10] (The Year 1 results were similar.)

The overall observation score averaged across four windows provided some reliable information to identify teachers who needed support, but single observations provided limited information on teachers' persistent performance. In Year 2, for example, depending on the assumptions used, reliability estimates for the four-window average overall scores were between .53 and .61 for the CLASS. This implies that 53 to 61 percent of the variation was due to persistent variation in the quality of teacher practice, and the rest (39 to 47 percent) was due to measurement error. Reliability estimates were between .70 and .77 for the FFT. Overall scores based on a single observation had limited reliability as a measure of a teacher's persistent classroom practice over each year because of variation in a teacher's overall scores across the four observation windows. In Year 2, the reliability of overall scores based on a single observation was .33 for CLASS and .51 for FFT.[11]

The observations provided limited information to signal specific areas of practice for improvement. While most teachers received ratings that differed across dimensions, the differences were not sufficiently reliable to identify dimensions for improvement, even when averaged over the year (.35 to .43 for the CLASS and .18 to .30 for the FFT in Year 2).

A majority of treatment teachers said the study's feedback on classroom practice was more useful and specific than the district's existing feedback. For example, about 65 percent of teachers reported that the study's feedback was more useful than their district's, and 79 percent reported that the study's feedback was more specific about what constitutes high-quality teaching.

How well was the student growth measure implemented and what information did the measure provide?

The student growth measure produced information on each teacher's contribution to student achievement using value-added methods. Value-added methods involve predicting the test score

[10] Teachers observed using the FFT instrument did not receive an overall score or overall performance level. For analytic purposes, the study's evaluation team created an overall score for the FFT by averaging the 10 FFT dimension scores and assigning this overall score to one of four study-defined performance levels.

[11] Classroom practice ratings from a single observation could also inform feedback about a teacher's instruction during a particular lesson, even if that performance were not indicative of a teacher's general instruction over the year. We do not have the necessary data to estimate the reliability of using single observations for feedback about instruction specific to a given lesson.

each student would have received, accounting for prior achievement and other characteristics, if the student had been taught by the average teacher in the district. A teacher's value-added score is obtained by comparing the average actual performance of the teacher's students to the average of the students' predicted scores.

Each year, value-added scores were generated for teachers of students in grades 4–8 reading/English language arts and mathematics classrooms in each district, using the achievement data for the students that each teacher had taught in the previous two years.[12,13] Each treatment teacher was given access to a "student growth" report that included the teacher's value-added scores along with an 80 percent confidence interval, which could be used to determine whether the scores were "measurably" different from the district's average.[14] Treatment principals were also given access to a report with their teachers' value-added scores and the school's average scores.

Fewer than half of teachers and principals accessed their growth reports in Year 1. In Year 2, almost all teachers received printed reports, and reports were viewed by all principals. In Year 1, despite good attendance at webinars encouraging educators to access their reports through an online portal (85 percent and 81 percent for teachers and principals, respectively), access rates were low—39 percent of the teachers with value-added scores and 40 percent of the principals.[15] To address this, in Year 2, each principal was given a printed school-level report and a packet for each teacher containing the teacher's most recent student growth report and classroom observation report; reports were viewed by all principals and were received by 98 percent of teachers.

Many teachers with a student growth report had value-added scores that measurably differed from the district average, particularly in mathematics, and the growth reports had the potential to signal which subject to focus on for improvement. In reading/English language arts, 23 percent of teachers in Year 1 and 21 percent in Year 2 had value-added scores that differed from the district average; in mathematics, 52 percent of teachers in Year 1 and 47 percent in Year 2 had value-added scores that differed from average.[16] Among teachers with value-added scores in both reading/English language arts

[12] A value-added score for a given subject was produced for a teacher only if the teacher had at least 10 students who had the necessary achievement data.

[13] In addition, student growth reports were prepared for teachers in Year 3, after the study was over, based on data in Years 1 and 2.

[14] The student growth reports used an 80 percent confidence interval (i.e., the range of scores that have an 80 percent chance of including the teacher's "true" score) to identify scores that were "measurably" below or above average. This benchmark was selected in order to appropriately balance the risk of misclassifying a teacher who was actually average as above or below average, against the risk of misclassifying a teacher who was actually above or below average as average. One consideration in striking this balance was that the study districts agreed that the value-added scores would not be used for decisions with consequences for employment. This reduced the potential downside associated with misidentifying an average teacher as below average.

[15] The analysis of teacher access rates was based on teachers with value-added scores. The analysis of principal access rates was based on all treatment schools in which at least one teacher had a value-added score. This included all but one school in the sample.

[16] The reliability estimates for teachers' value-added scores were 0.44 for reading/English language arts and 0.68 for mathematics in Year 1, and 0.46 and 0.67, respectively, in Year 2.

and mathematics, about half had student growth reports that suggested the teacher performed better in one subject area than the other, potentially identifying an area for improvement.

How well was the principal leadership measure implemented and what information did the measure provide?

The third component of the intervention was intended to provide feedback on multiple dimensions of the principal's effectiveness as a leader. This feedback was based on the Vanderbilt Assessment of Leadership in Education (VAL-ED), a 360-degree survey assessment administered twice a year to principals, principal supervisors, and teachers. A report for each principal was generated after each administration of the VAL-ED, which the principal was to discuss with his or her supervisor in a feedback session. The report included ratings on dimensions of leadership, as well as an overall score and performance levels (*below basic*, *basic*, *proficient*, *distinguished*).

Principal feedback sessions generally occurred as planned. After each VAL-ED administration, nearly all principals met with their supervisors to discuss their reports.[17] In Year 1, principals' supervisors reported that the feedback sessions lasted 52 minutes on average in the fall and 46 minutes in the spring. In Year 2, the sessions lasted 36 minutes in the fall and 34 minutes in the spring.

In all four administrations, principals' scores were distributed across all four VAL-ED performance levels, and thus many principals received scores indicating a need for improvement. In the fall of Year 1, 70 percent of principals were in the bottom two performance levels. In the spring of Year 2, 41 percent were in the bottom two levels.

The VAL-ED ratings provided by principals, their supervisors, and the teachers in their schools were often too different from each other to form a reliable measure in the fall administrations, but the spring ratings were consistent enough to identify principals who needed support. Based on the literature on 360-degree surveys, we would expect correlations of 0.25 to 0.35 between respondent group scores.[18] In the fall administrations, however, agreement among the three groups' overall scores was low, with two of the three correlations below 0.10. In the spring, correlations were higher (0.23 to 0.38), providing a more reliable message about a principal's effectiveness. Almost all reports showed dimension scores that spanned multiple performance levels, but these scores did not reliably indicate which dimension a principal most needed to work on.

Nearly three-quarters of treatment principals reported that the study's feedback on their leadership was more objective and actionable than previous feedback from their district. For example, 73 percent of treatment principals reported that the VAL-ED feedback was more objective than feedback they had previously received from their districts, and

[17] In each of the two study years, each principal in a treatment school participated in at least one feedback session. In Year 2, a small number of principals did not participate in a second feedback session.

[18] For the VAL-ED correlations, see Porter et al. (2010). For the literature on 360-degree surveys, see Conway and Huffcutt (1997).

75 percent reported that the VAL-ED feedback provided “clearer ideas about how to improve my leadership.”

Contrast in Educators’ Experience of Feedback

The study’s performance feedback was provided in addition to the districts’ established teacher and principal evaluation systems. It was intended to increase the frequency of feedback and to incorporate numerical ratings and, for teachers, a written narrative justification.

Did the intervention increase feedback for teachers?

As expected, treatment teachers reported receiving more feedback than control teachers. Each year, more than 80 percent of treatment teachers reported receiving feedback that included numerical ratings, compared with fewer than half of the control teachers.[19] Each year, treatment teachers also reported more than four times as many feedback sessions with ratings and a written narrative on their classroom practice as control teachers did. In both years, the average treatment teacher reported 3.0 feedback sessions that included ratings and a written narrative, compared with 0.7 for the average control teacher in Year 1 and 0.2 instances in Year 2. (See exhibit ES.1.) The total length of all feedback sessions was also substantially larger for treatment than control teachers—for example, 100 minutes in Year 2 for the average treatment teacher, compared with 25 minutes for the average control teacher.

[19] The data on feedback are based on a survey administered in the spring of each year, which asked teachers to report on every instance in which they were observed and later received feedback that year, including evaluation-related observations as well as walkthroughs and informal observations (e.g., peer-to-peer observations).

Exhibit ES.1. Number of feedback sessions with ratings and written narrative and duration of oral feedback that an average teacher reported receiving, by treatment status and year

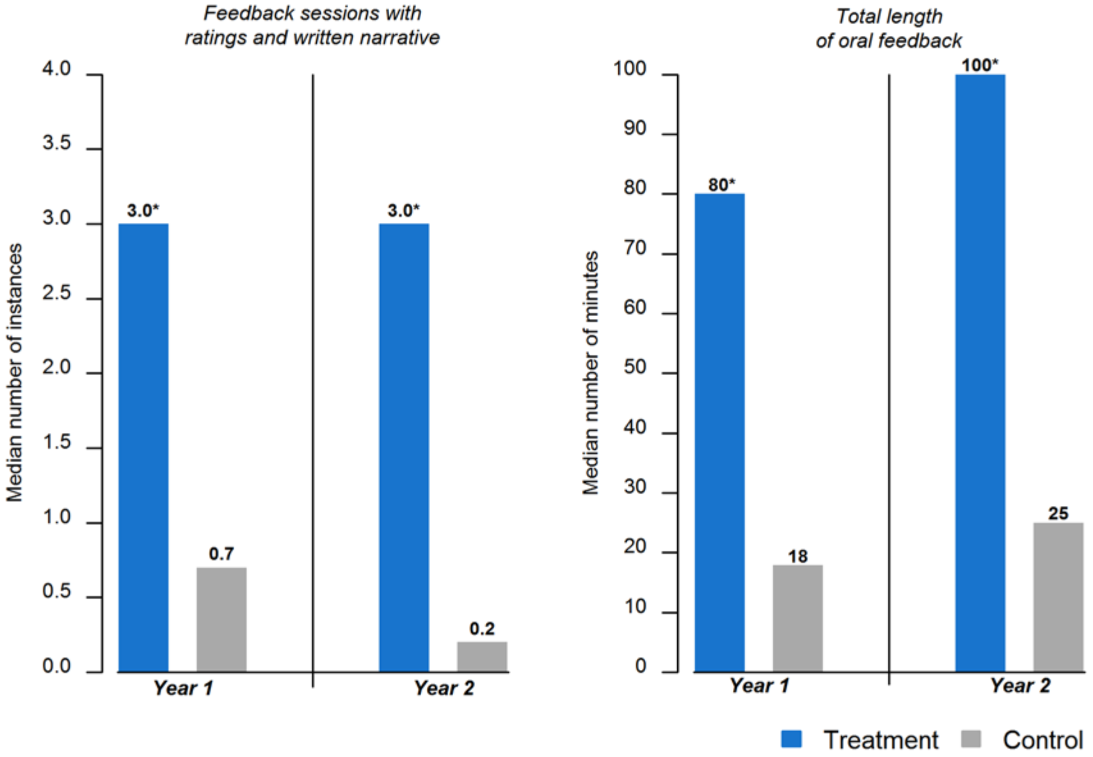

EXHIBIT READS: The average treatment teacher in Year 1 reported 3.0 feedback sessions with ratings and written narrative, compared with 0.7 for control teachers.

NOTES: Year 1 sample size = 63 schools and 523 teachers for the treatment group; 64 schools and 549 teachers for the control group. Year 2 sample size = 63 schools and 495 teachers for the treatment group; 63 schools and 521 teachers for the control group.

The analyses were based on an aligned rank sum test with randomization inference about median difference between treatment and control groups (see appendix H for technical details).

* Difference between the treatment and control groups is statistically significant at the .05 level (two-tailed).

SOURCES: Spring 2013 and Spring 2014 Teacher Surveys.

Treatment teachers were also more likely than control teachers to report receiving value-added scores. In Year 1, 45 percent of treatment teachers reported receiving value-added scores, compared with 24 percent of control teachers; in Year 2, the numbers were 81 and 34 percent, respectively.[20]

Did the intervention increase feedback for principals?

In both years, treatment principals reported receiving more feedback with ratings than control principals. Treatment principals reported receiving more instances of oral

[20] The survey items asking teachers whether they received value-added information differed in Years 1 and 2. In Year 1, the item was included in a broader question asking about different types of achievement information. In Year 2, the survey included a separate question asking whether teachers viewed a value-added score representing the classes they taught.

feedback with ratings than control principals (1.0 versus 0.4 instances in Year 1, and 2.0 versus 1.0 instances in Year 2).[21] (See exhibit ES.2.) In addition, as expected, in both years, the average treatment principal reported receiving a larger amount of oral feedback than did the average control principal (60 versus 41 minutes in Year 1, and 60 versus 33 minutes in Year 2).

Exhibit ES.2. Number of feedback instances and duration of oral feedback that principals reported receiving, by treatment status and year

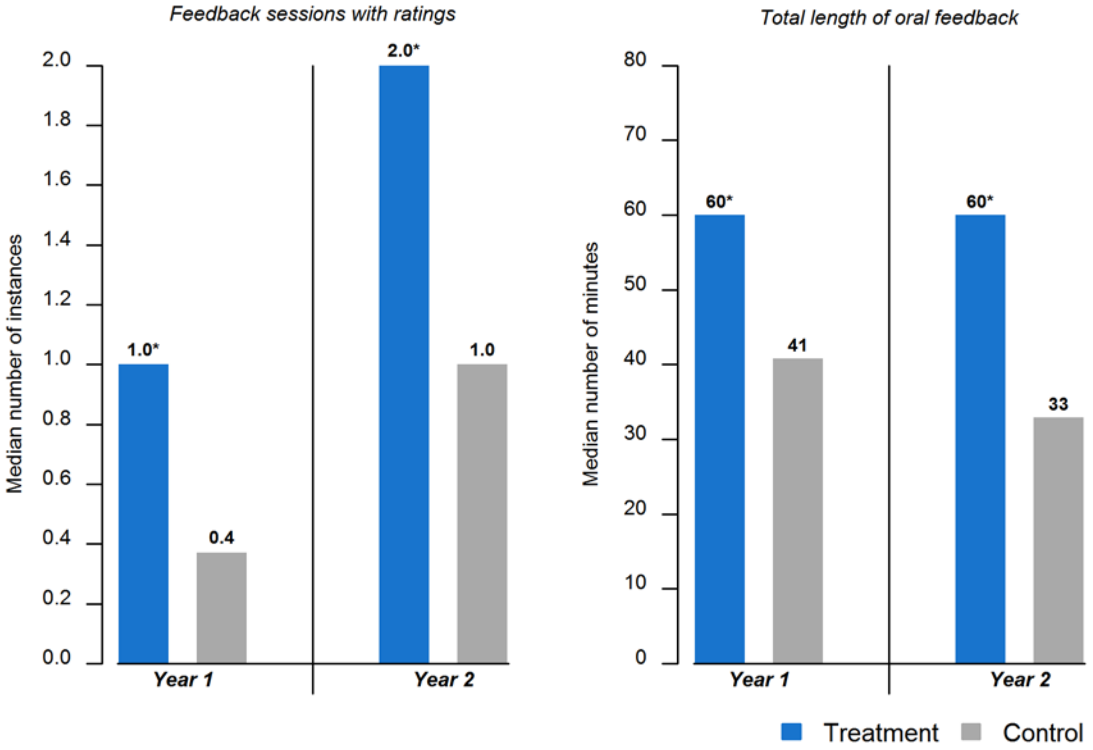

EXHIBIT READS: The average treatment principal in Year 1 reported receiving 1.0 feedback sessions with ratings, compared with 0.4 for control teachers.

NOTES: Year 1 sample size = 61 treatment and 61 control principals. Year 2 sample size = 61 treatment and 59 control principals. The analyses were based on an aligned rank sum test with randomization inference about median difference between treatment and control groups (see appendix H for technical details).

* Difference between the treatment and control groups is statistically significant at the .05 level (two-tailed).

SOURCES: Spring 2013 and Spring 2014 Principal Surveys.

Impact on Classroom Practice, Principal Leadership, and Student Achievement

The main premise behind providing performance feedback is that it would improve teachers' classroom practice and principals' leadership, and ultimately student achievement. Impacts on these outcomes could occur in at least two ways. First, feedback could influence whether more-effective teachers and principals remained in their schools, and whether less-effective staff left and were replaced by more-effective staff. Second, feedback could improve the practice of

[21] The principal survey was administered later in the spring in Year 2 than in Year 1, permitting the principals to include feedback that occurred later in the school year. This may explain why both treatment and control principals reported more instances of feedback in Year 2 than in Year 1.

teachers and principals who stayed. The analyses we conducted focused on all teachers and principals present in the study schools in the spring of Years 1 and 2, and thus the sample included some educators who stayed and some who were new to their schools. Any impacts observed thus reflect a mix of effects on educator mobility and on improvement of those who stayed.

Did the intervention have an impact on classroom practice?

To provide a common outcome measure to use in assessing the impact on teacher classroom practice, we video-recorded one lesson for each treatment and control teacher in the spring of Year 2 and a second lesson for a random sample of half the teachers. Each lesson was coded by trained observers using *both* the CLASS and the FFT instruments. We used both instruments so we could assess whether the feedback had an impact on the practices measured by the instrument on which the feedback was based, and also on an instrument that measured practices that were similar but not exactly those used as a basis for the feedback.

The intervention had a positive impact on teachers' classroom practice based on video-recorded lessons coded using the CLASS, but not on practice coded using the FFT. On average, treatment teachers received a score of 4.50 on the CLASS (on the 7-point CLASS scale), compared with 4.39 for control teachers. (See exhibit ES.3.) The 0.11-point difference corresponds to an improvement index of 7 percentile points, implying that the percentile rank of the average control teacher would increase from the 50th percentile to the 57th percentile if the teacher received the intervention. There was no statistically significant difference between the treatment and control teachers when classroom practice was coded using the FFT.

We also estimated the impact on classroom practice as measured by video-recorded lessons separately for the four districts that used CLASS for feedback and the four that used FFT, anticipating that, at a minimum, there might be an impact on the aligned practice measures (i.e., an impact on CLASS scores in districts that used the CLASS for feedback, and an impact on FFT scores in districts that used the FFT for feedback). We found a 0.31-point impact on CLASS scores in the four CLASS districts (corresponding to an improvement index of 18 percentile points). There was no statistically significant impact on CLASS scores in the FFT districts, however, and there was no impact on FFT scores in either CLASS or FFT districts. Because study districts chose to use the CLASS or the FFT as part of the intervention, we cannot draw definitive conclusions about why an impact on classroom practice was found in CLASS but not in FFT districts.

Exhibit ES.3. Average CLASS and FFT scores, based on coding of video-recorded lessons by study team, by treatment status, Year 2

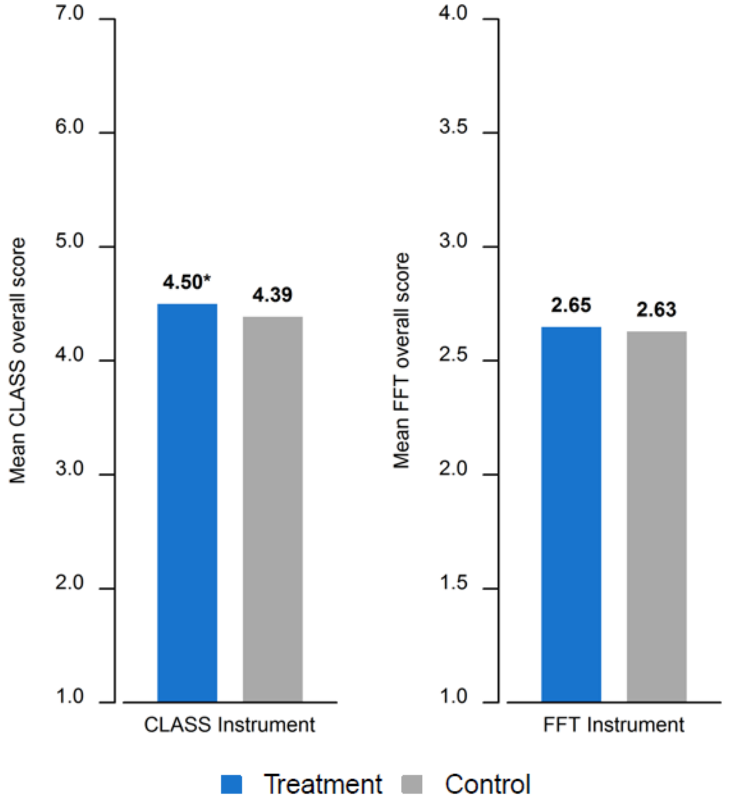

EXHIBIT READS: The average CLASS overall score was 4.50 for treatment teachers, compared with 4.39 for control teachers.

NOTES: Sample size = 63 schools, 434 teachers, and 668 videos for the treatment group; 63 schools, 517 teachers, and 793 videos for the control group. The analyses were based on a three-level regression (lessons within teachers within schools) controlling for random assignment blocks and teacher background characteristics.

* Difference between the treatment and control groups is statistically significant at the .05 level (two-tailed).

SOURCE: Spring 2014 Classroom Videos.

Did the intervention have an impact on principal leadership?

The goal of the principal feedback was to improve their leadership skills. We measured two aspects of leadership: instructional leadership and teacher-principal trust.

The intervention had a positive impact on teacher-principal trust in Year 1 and on both instructional leadership and teacher-principal trust in Year 2. In Year 1, treatment principals, on average, received a score of 3.18 on the 5-point teacher-principal trust scale, compared with 2.96 for control principals. (See exhibit ES.4.) The 0.22-point difference corresponds to an improvement index of 10 percentile points, implying that the trust score for the average control principal would increase from the 50th percentile to the 60th percentile if the school received the intervention. In Year 2, there were positive impacts on both instructional leadership (0.14 points) and teacher-principal trust (0.15 points). Although there were statistically significant impacts on both leadership measures in Year 2, and only one in Year 1, the magnitudes of the impacts did not statistically differ in the two years, and thus there is little evidence for an increase in impact over the two years.

Exhibit ES.4. Average rating of principal instructional leadership and teacher-principal trust, by treatment status and year

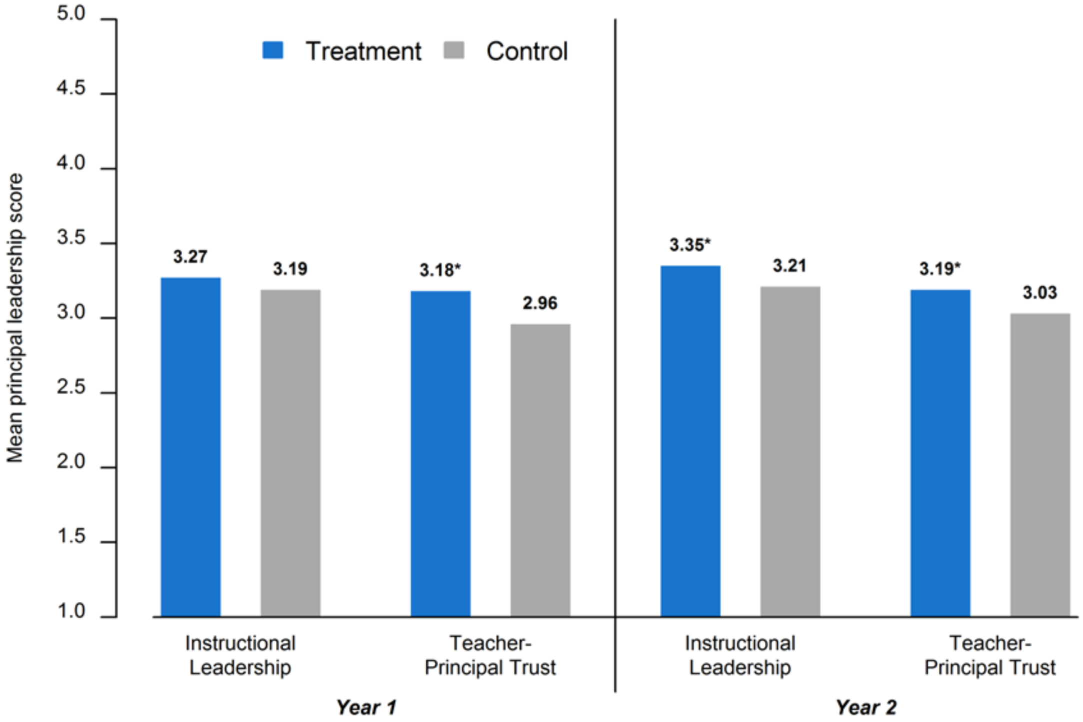

EXHIBIT READS: The average rating of principals' instructional leadership in treatment schools in Year 1 was 3.3, compared to 3.2 for principals in control schools.

NOTES: Year 1 sample size = 63 principals and 524 or 525 teachers for the treatment group; 64 principals and 557 teachers for the control group. Year 2 sample size = 63 principals and 499 teachers for the treatment group; 63 principals and 522 or 523 teachers for the control group. The analyses were based on a two-level regression (teachers nested in schools) controlling for random assignment blocks and teacher background characteristics.

* Difference between the treatment and control groups is statistically significant at the .05 level (two-tailed).

SOURCES: Spring 2013 and 2014 Teacher Surveys.

Did the intervention have an impact on student achievement?

The ultimate goal of the intervention was to boost students' achievement in reading/English language arts and mathematics. We examined the impact on achievement by comparing students' scores on the state achievement test for all students enrolled in treatment and control teachers' classes in the spring of Year 1 and in the spring of Year 2. The Year 1 estimates controlled for student achievement in the spring of the year before the intervention was implemented (i.e., the baseline year), and thus the estimates represent the effect of the first year of implementation of the intervention. The Year 2 estimates also controlled for student achievement from the baseline year, and thus they represent the cumulative impact of the intervention over two years.

The intervention had a positive impact on students' mathematics achievement in Year 1, and had a cumulative impact similar in magnitude but not statistically significant (*p* = 0.055) in Year 2. The intervention did not have an impact on students' reading/English language arts achievement in either year. In Year 1, in mathematics, students in treatment schools scored at the 51.8th percentile in their district, compared to the 49.7th percentile for control students. (See exhibit ES.5.) The 2.1-point difference corresponds to about one month of learning.[22] In Year 2, in mathematics, students in treatment schools scored at the 51.2nd percentile, compared to the 48.9th percentile for control students, a 2.3-point difference, similar in magnitude to the impact in Year 1 but not statistically significant ($p = 0.055$). The impacts for reading/English language arts (0.4 points in Year 1 and 1.0 in Year 2) were smaller than the impacts for mathematics and were not statistically significant. There is no evidence that the cumulative impact on achievement increased from the first to the second year of implementation.

Exhibit ES.5. Average reading/English language arts and mathematics achievement, by treatment status and year

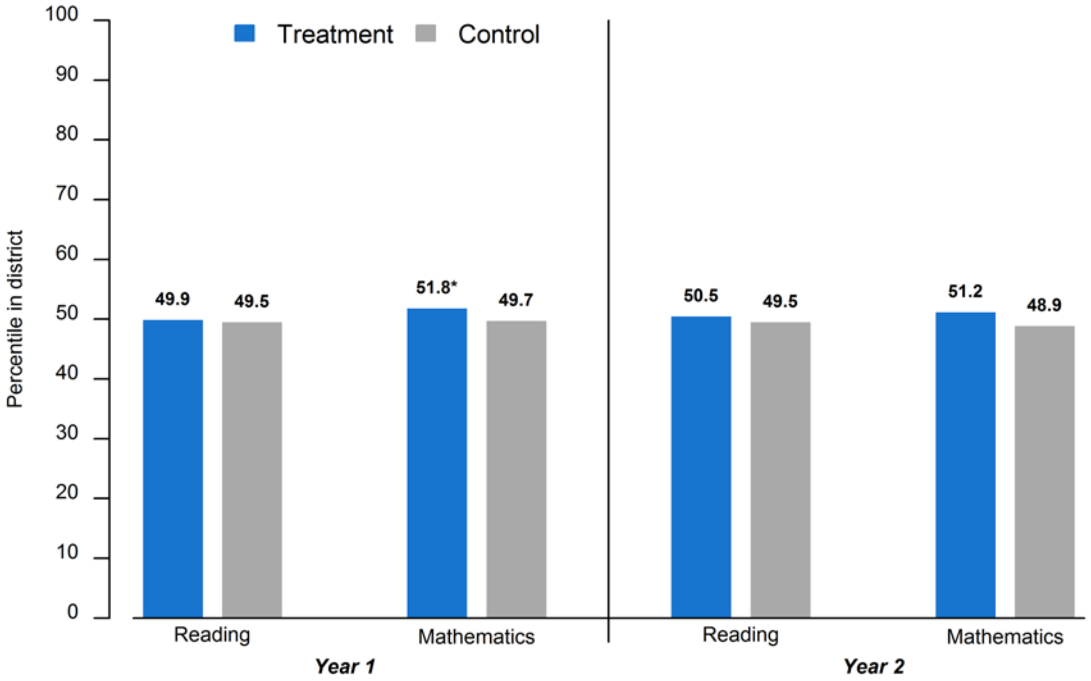

EXHIBIT READS: In Year 1, students in treatment schools received an average reading/English language arts score at the 49.9th percentile in their district, compared to the 49.5th percentile for students in control schools.

NOTES: Sample size for Year 1 reading/English language arts = 63 schools, 384 teachers, and 13,134 students for the treatment group; 64 schools, 421 teachers, and 15,358 students for the control group. Sample size for Year 1 mathematics = 63 schools, 411 teachers, and 13,967 students for the treatment group; 64 schools, 439 teachers, and 15,907 students for the control group. Sample size for Year 2 reading/English language arts = 63 schools, 374 teachers, and 13,962 students for the treatment group; 63 schools, 394 teachers, and 15,423 students for the control group. Sample size for Year 2 mathematics = 63 schools, 389 teachers, and 14,186 students for the treatment group; 63 schools, 396 teachers, and 15,809 students for the control group. The analyses were based on a three-level regression (students nested within teachers within schools) controlling for random assignment blocks and student background characteristics.

* Difference between the treatment and control groups is statistically significant at the .05 level (two-tailed).

SOURCE: District Administrative Records.

[22] According to Hill et al. (2008), the average annual gain in mathematics is about 0.42 standard deviations for students in grades 4–8. The impact of 2.10 percentile points is about 0.05 standard deviations. This translates into about 0.05/0.42 = 0.11 of a year's achievement gain. Assuming a 36-week school year, this implies that the impact corresponds to four weeks of learning.

Association Among Classroom Practice, Leadership, and Achievement

The study's theory of action assumed that performance feedback for educators would improve student achievement by improving teachers' practice and principals' leadership. The study was not designed to provide a rigorous causal test of this assumption. However, exploratory analyses indicate that classroom practice, using the study's outcome measure based on video-recorded lessons coded with the CLASS and the FFT, was positively associated with student achievement in mathematics and reading, suggesting that improved classroom practice may have been one way feedback boosted achievement.[23] Similar exploratory analyses found no association between the study measures of leadership and achievement.

[23] We examined whether teachers' classroom practice based on video-recorded lessons was associated with their students' reading and mathematics achievement, controlling for students' prior achievement and other student and teacher background characteristics. We found an association with students' mathematics achievement of 0.06 for classroom practice as measured by the CLASS and 0.07 as measured by the FFT. We found an association with students' reading achievement of 0.03 for classroom practice as measured by the CLASS and also as measured by the FFT.

This page has been left blank for double-sided copying