U.S. Department of Education
November 2016

Early Implementation Findings From a Study of Teacher and Principal Performance Measurement and Feedback: Year 1 Report

Andrew J. Wayne
Michael S. Garet
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.

Early Implementation Findings From a Study of Teacher and Principal Performance Measurement and Feedback: Year 1 Report

November 2016

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

Melanie Ali
Project Officer
Institute of Education Sciences

NCEE 2017-4004
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
John King
Secretary

Institute of Education Sciences
Ruth Curran Neild
Deputy Director for Policy and Research
Delegated Duties of the Director

National Center for Education Evaluation and Regional Assistance
Joy Lesnick
Acting Commissioner

November 2016

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:

Wayne, A. J., Garet, M. S., Brown, S., Rickles, J., Song, M., Manzeske, D., (2016). *Early Implementation Findings From a Study of Teacher and Principal Performance Measurement and Feedback: Year 1 Report* (NCEE 2017-4004). 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.

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

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, Nicole 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, Sara 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. Finally, we are grateful to Connie Conroy who provided administrative assistance throughout the project.

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 Early Implementation Findings From a Study of Teacher and Principal Performance Measurement and Feedback. 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.

Contents

Page

List of Exhibits

Page

Executive Summary

Educator performance evaluation systems are a potential tool for improving student achievement.[1] By removing ineffective teachers and principals and/or through increasing the effectiveness of the existing workforce, such systems may result in higher student achievement.[2]

Emerging research suggests some promising features of performance evaluation measures. For example, research suggests that, to measure classroom practice, additional observations of the same teacher, beyond the first, form a more reliable measure of a teacher's typical practice, especially when more than one observer is used.[3] There is also some evidence from recent research that giving more frequent, specific feedback on classroom practice may lead to improvements in teacher performance and student achievement.[4]

The U.S. Department of Education's Institute of Education Sciences is conducting a study on the implementation and impacts of teacher and principal performance measures that are consistent with emerging research. 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:

- a measure of teacher classroom practice with subsequent feedback sessions conducted four times per year, based on a classroom observation rubric;
- a measure of teacher contributions to student achievement growth (i.e., value-added scores), provided to teachers and their principals once per year; and
- a measure of principal leadership with subsequent feedback sessions conducted twice per year, using a leadership survey administered to the principal, the principal's supervisor, and the principal's teachers.[5]

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.[6] Instead, the measures

[1] See Stecher et al. (2016); Weisburg, Daniel, Sexton, Mulhern, and Keeling, (2009).

[2] Researchers studying a range of educator workforce interventions theorize that there are two key mechanisms leading to improved student achievement: changes in the composition of the workforce and changes in the skills of continuing workers. See, for example, Taylor and Tyler (2012) regarding teacher evaluation, Chiang et al. (2015) regarding performance-based compensation, and Glazerman et al. (2010) regarding comprehensive teacher induction.

[3] See Bill & Melinda Gates Foundation (2012); Whitehurst, Chingos, and Lindquist (2014).

[4] See Steinberg and Sartain (In Press); Taylor and Tyler (2012).

[5] These measures are the kinds of measures emphasized through the Elementary and Secondary Education Act flexibility waivers, as well as federal programs such as Race to the Top and the Teacher Incentive Fund. They also are the kinds of measures that states are allowed to develop and support with Title II, Part A funds under the Every Student Succeeds Act of 2015.

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

were used to provide educators and their supervisors with information regarding performance. Such information might affect perceptions about performance, motivate improvement, and/or lead to improved knowledge and skills among educators.

This report focuses on the first year of the two years of implementation, describing the characteristics of the educator performance measures and teachers' and principals' experiences with performance feedback. It is one of the few detailed descriptions of educator performance measures implemented on a large scale in districts.[7] The final report will primarily focus on impacts on outcomes including principal leadership, teacher classroom practice, and student achievement but will also include descriptive information on the second year of implementation.

The main findings in this report are as follows:

- **Educator performance measures were implemented generally as planned, except that fewer than the intended number of educators accessed the student growth reports.** Most teachers received the intended four rounds of observations and feedback sessions (mean = 3.8 observations and 3.7 feedback sessions per teacher) by trained and certified observers. Only 40 percent of principals and 39 percent of teachers with value-added scores accessed their student growth reports. All principals in the implementing schools received two reports about their leadership practices based on their survey responses and that of their teachers and their supervisor. They also met with their supervisors to discuss their reports, in both fall and spring.
- **Both classroom observation and student growth measures differentiated teacher performance, but observation scores were skewed toward the upper end of the scale.** In each classroom observation window, a large majority of the teachers observed had classroom observation overall scores in the top two performance levels (more than 85 percent), and very few teachers had overall scores in the lower two levels (less than 15 percent). However, there was some variation in scores across teachers, and both value-added scores and average classroom observation scores over the year had sufficient reliability to capture performance differences among some teachers. About one quarter of the teachers with value-added scores in reading/English language arts and one half of the teachers with value-added scores in mathematics had student growth reports indicating that their score in that subject was measurably below or above the district average.[8]
- **The principal leadership measure differentiated performance, but there was limited consistency in scores across survey respondent groups.** Principals'

[7] For an example, see Lipscomb, Terziev, and Chaplin (2015).

[8] To test whether teachers' value-added scores differed from the district average, we used an 80 percent confidence interval. There were two reasons for the use of an 80 percent rather than a 95 percent confidence interval more typical of statistical tests. First, the student growth report available to principals and teachers in the intervention schools included the score with an 80 percent confidence interval. As part of the student growth report training, educators were told to use this confidence interval to determine if their score differed from average performance in the district. Second, the value-added scores were used for informational purposes and not decisions with consequences for employment, for example. Given this, the 80 percent confidence interval was intended to balance the risks associated with mistakenly classifying average performers as above or below average with the risks associated with mistakenly classifying above or below average performers as average.

overall leadership scores were spread across the four performance levels. Half or more of principals in the schools selected to implement the intervention received overall scores on the principal leadership measure in the lower two categories labeled *basic* or *below basic* (70 percent in the fall and 51 percent in the spring administration). However, the respondent groups (principals, teachers, and principal supervisors) often scored principals differently.

- **Both teachers and principals in schools selected to implement the intervention reported receiving more feedback than those in schools in the same districts selected to continue with business-as-usual.** For example, teachers in intervention schools reported more feedback sessions that were accompanied by a rating and written narrative (3.0 versus 0.7 sessions).[9] They also reported spending more total time in feedback sessions (80 minutes versus 18 minutes). Principals in intervention schools reported more instances of receiving feedback with ratings (1.0 versus 0.4) and spending more total time in feedback sessions (60 minutes versus 41 minutes).

Study Overview

The purpose of this study is to describe teachers' and principals' experiences with the study's performance measures and feedback over two years, and to examine whether the information provided by the measures and feedback affected educator and student outcomes. The study has five research questions:

1. To what extent were the performance measures and feedback implemented as planned?
2. To what extent did the performance measures distinguish educator performance?
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?

This report addresses the first three questions, focusing on the first year of implementation. A subsequent report will address the first three questions focusing on the second year of implementation. That report will also address the last two questions.

Study Design

The study examines the implementation and impacts of an intervention consisting of three performance measures with feedback for teachers and principals through an experimental design conducted 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 systematic and intensive than that implemented as part of the study.

[9] Throughout this executive summary, all treatment-control differences noted are statistically significant using two-tailed tests and applying a 5 percent critical value, unless noted otherwise.

Recruited districts required fewer than four observations of teachers per year. In addition, the districts' evaluation systems did not require the inclusion of student achievement information in teacher ratings. None of the recruited districts used a leadership measure similar to that used by the study.

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

Each study district identified a set of regular elementary and middle schools that were willing to participate in the study. In these schools, the study focuses on the teachers of mathematics and reading/English language arts in grades 4-8, as well as the principals.[10] The schools were assigned by lottery to implement the three measures with feedback (the treatment group) or not (the control group). Both groups continued to implement their district's existing performance evaluations and measures, and the treatment group 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 are 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 for this report came from multiple sources as described next.

Data on the implementation of the intervention. We documented attendance at orientation and training events related to the study's performance measures. Online system records maintained by the vendors of the measures were used for information on observer certification test pass rates, the frequency and timing of teacher observations and feedback sessions, and teachers' and principals' accessing of student growth reports. Surveys of observers hired by the study and interviews with district officials provided further information regarding the implementation of the observations and the district context, respectively.

Data on measures of educator performance. Data on measures of teacher classroom practice, student growth, and principal leadership were collected through the vendors' online systems.

Data on educators' experiences with performance feedback. In spring 2013, we surveyed both the principals and teachers in all treatment and control schools. These surveys

[10] 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 are not included in the main study analyses, however, because student assessment data are not available in Kindergarten through grade 3.

collected information on the nature and frequency of performance information educators received and their perceptions of that information.

Data on the characteristics of study participants. To compare the characteristics of participants in the treatment and control groups, we collected data on school characteristics from the 2011–12 Common Core of Data and collected data on principals', teachers', and students' characteristics from district administrative records.

Analyses

To examine the implementation of the performance measures, we describe the extent to which study participants received the training on the measures, carried out the performance measurement activities, and received performance information and feedback as planned. We also examined the characteristics of the ratings teachers and principals received, including whether they distinguish between lower and higher performers. These analyses yielded the average rating scores, the percentage of ratings in each performance level, and the variation in the ratings across teachers and across principals. To assess whether the study's intervention led to differences in educators' experiences with performance measurement and feedback, we compared survey responses of teachers and principals in the treatment and control groups.

Detailed Summary of Findings

The following section provides additional information about the extent to which each of the study measures was carried out as intended and whether the information from the measures distinguished between lower- and higher-performing educators and thus could be used to identify educators in need of support. These analyses pertain only to teachers and principals in the treatment schools. This section also highlights the extent to which educators' experiences with performance information differed between the treatment and control groups.

The Classroom Practice Measure and Feedback

The teacher classroom practice measure was based on four classroom observations during the school year. For each teacher, one observation was to be conducted by a school administrator and the other three by observers hired by the study. After each observation, the observer was expected to prepare a standard report with both ratings and narrative justification and to discuss the report with the teacher during a feedback session. Both teachers and their principal had access to the standard report.

How Many Observations Were Conducted and What Were Observers' Qualifications?

- **Observers were trained and certified as planned.** Nearly all observers (92 percent for CLASS and 97 percent for FFT) completed all of the required training, which lasted three days for CLASS and four days for FFT. All observers passed the certification test, although it took multiple attempts to pass the test for half of the CLASS trainees and 17 percent of the FFT trainees.

- **The majority of teachers were observed the intended four times and received feedback.** The majority of teachers (73 percent for CLASS and 95 percent for FFT) received all four observations, and the majority of teachers (57 percent for CLASS and 94 percent for FFT) also received all four feedback sessions, as intended. On average, teachers received 3.8 observations (3.7 for CLASS and 3.9 for FFT) and 3.7 feedback sessions (3.5 for CLASS and 3.9 for FFT) during the first year of the study.

What Were the Characteristics of the Classroom Practice Performance Information Provided?

- **CLASS reports provided separate scores for individual dimensions as well as the teacher's overall score and a sense of how their performance compared with others; FFT reports provided only separate scores for individual dimensions.** The CLASS reports included scores for 12 dimensions of teaching grouped into four teaching domains, as well as an overall score for the observation and a score for each domain (emotional support, classroom organization, instructional support, and student engagement). In addition, the CLASS reports included comparisons with the district average scores and the teacher's prior scores. The FFT reports provided scores for up to 10 dimensions of teaching grouped into two teaching domains (classroom environment and instruction). The FFT reports did not include an overall score, domain scores, scores from past observations, or district average scores.
- **Most of the CLASS observation reports identified at least one dimension of classroom practice to improve and illustrated it with an example from the observation, but less than a quarter of FFT reports did so.** The observers were required to write narrative text identifying at least one dimension of practice as a strength and one dimension for improvement. The majority of the observation reports (76 percent of CLASS reports and 71 percent of FFT reports) did so. In addition, three quarters of the CLASS reports supported the identified dimension(s) for improvement with at least one example from the observation, but less than a quarter (23 percent) of the FFT reports did so.[11]
- **For both CLASS and FFT, observation scores were concentrated at the upper end of the scale, limiting the degree of differentiation between lower- and higher-performing teachers.** Nearly all teachers had CLASS or FFT overall scores for a given observation window in the top two performance levels (more than 95 percent of the CLASS scores and more than 85 percent of the FFT scores). Only a small percentage of the teachers had scores consistent with the lowest two performance levels (under 5 percent for CLASS and under 15 percent for FFT depending on the observation window).[12] (See exhibits ES.1 and ES.2.) While most teachers had overall scores in the top two performance levels, many teachers had dimension-level scores at different performance levels (e.g., in the first window 61 percent of CLASS teachers and 69 percent of FFT teachers received scores at multiple performance levels).

[11] The findings reported here are based on an analysis of 160 randomly selected reports.

[12] Teachers observed using the FFT instrument did not receive an overall score or overall performance level for each observation window. For analytic purposes, the study's evaluation team calculated each teacher's average score in each observation window based on the 1 to 4 rating for each dimension of practice.

- **Teachers' overall classroom observation scores, averaged across all four windows, contained measurement error, but provided some reliable information to distinguish between lower- and higher-performing teachers and were positively correlated with teacher value-added scores.** Classroom observation scores averaged across the four observation windows had some reliability to help distinguish average teacher performance (reliability estimated between .42 and .50 for CLASS and .69 and .75 for FFT). These estimates, while lower than conventional thresholds for measures used in research, are consistent with findings from other studies of classroom observation reliability.[13] In addition, the CLASS and FFT four-window average scores were positively, although weakly, associated with teachers' prior-year value-added scores (correlations of .09 and .17, respectively).[14]
- **Differences in a teacher's ratings across observations limited how much one could learn about persistent performance from a single observation.** Less than half of the variation in teacher scores from a given observation window reflected stable classroom practice over the year. The reliability estimate for a single observation was .24 for CLASS scores and .49 for FFT scores, which indicates that 24 percent of the variation in CLASS scores and 49 percent of the variation in FFT scores reflected stable practice over the year.

[13] See Casabianca et al. (2013); Ho and Kane (2013); Kane and Staiger (2012).

[14] Although the correlations between classroom observation overall scores and value-added scores were modest in magnitude, these correlations are consistent with the magnitudes found by other studies (Chaplin et al. 2014; Kane and Staiger 2012; Kane et al. 2011) and likely underestimate the strength of the true association because of measurement error in both the observation scores and the value-added scores.

Exhibit ES.1. Distribution of treatment teachers across performance levels based on CLASS overall scores, by observation window

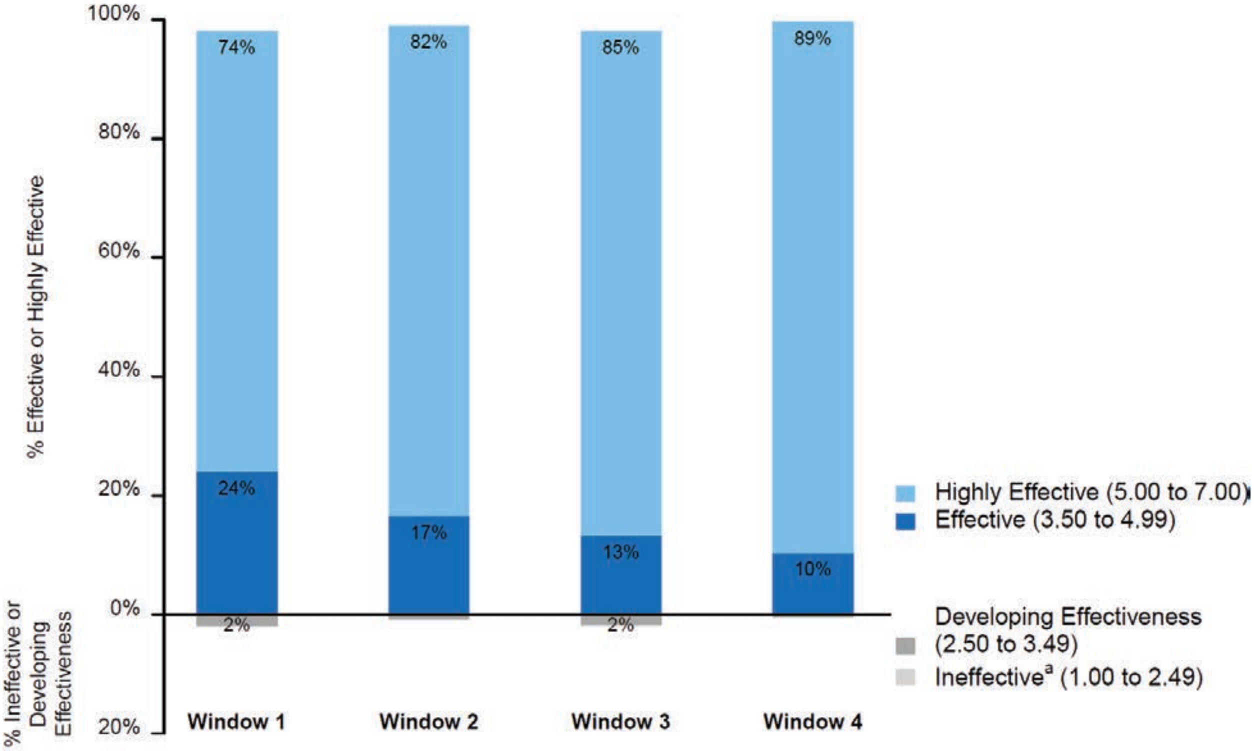


Exhibit Reads: Of treatment teachers in CLASS districts observed in window 1, 74 percent had a CLASS overall score at the *highly effective* performance level, 24 percent at the *effective* performance level, and 2 percent at the *developing effectiveness* performance level. Less than 1 percent of teachers had an overall score at the *ineffective* performance level.

NOTE: Performance level distributions are based on teachers' overall CLASS ratings in each window. Sample size = 262 teachers in window 1, 307 teachers in window 2, 309 teachers in window 3, and 272 teachers in window 4. Reported percentages may not sum to 100 percent because of rounding.

[a] Within a window, less than 1 percent of teachers had an overall score at the *ineffective* performance level.

SOURCE: Teachstone Online System.

Exhibit ES.2. Distribution of treatment teachers across study-defined performance levels based on FFT overall scores, by observation window

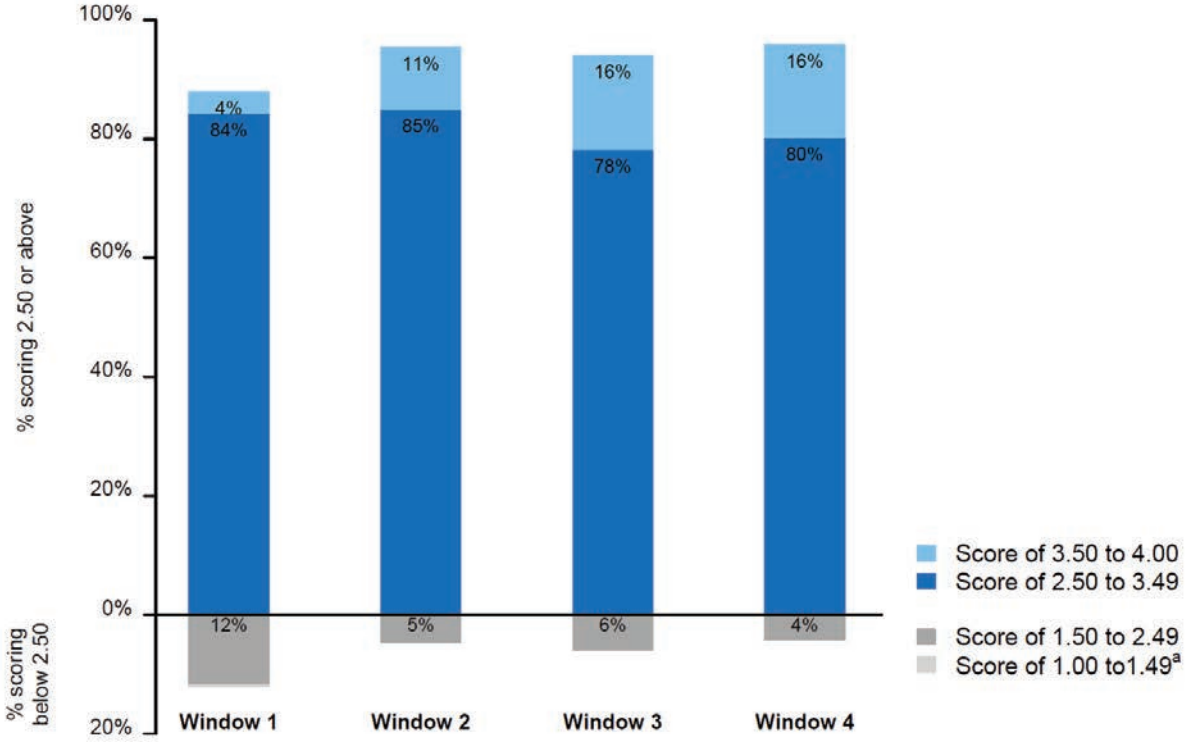


Exhibit Reads: Of treatment teachers in FFT districts observed in window 1, 4 percent had an FFT overall score between 3.50 and 4.00, 84 percent had a score between 2.50 and 3.49, and 12 percent had a score between 1.50 and 2.49. Less than one percent of teachers had an overall score below 1.50.

NOTE: The distribution in each window is based on teachers' FFT overall scores categorized into study-defined performance levels. To create the overall scores and performance levels, the study's evaluation team first calculated an overall score by averaging the teacher's ten FFT dimension scores, each of which was rated on a 1 to 4 scale. The overall scores were then categorized into study-defined performance levels by rounding them to the nearest whole number. This created four performance levels aligned with the FFT dimension scores. An FFT dimension score of 1 corresponds to *unsatisfactory*, 2 corresponds to *basic*, 3 corresponds to *proficient*, and 4 corresponds to *distinguished*. Average FFT scores and overall performance levels were not provided in the FFT reports teachers received. Sample size = 216 teachers in window 1, 219 teachers in window 2, 220 teachers in window 3, and 217 teachers in window 4. Reported percentages may not sum to 100 percent because of rounding.

[a] Within a window, less than 1 percent of teachers had an overall score below 1.50.

SOURCE: Teachscape Online System.

The Student Growth Measure

The measure of student growth was designed to provide teachers with information on their contribution to student achievement, using value-added methods. Value added methods involve predicting the test score 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.

Teacher value-added scores were generated for all teachers of students in grades 4–8 reading/English language arts and mathematics in each district using the achievement data for the students that each teacher taught in the previous two years.[15] Individual teachers in the treatment schools in these grades were given access to a report on their scores during the first year of implementation. Treatment principals were also given access to a report that included their teachers' student growth reports as well as school average value-added scores, overall and by subject and grade.

Who Received the Student Growth Performance Information?

- **A large majority of teachers had sufficient data to produce student growth reports.** Overall, student achievement data were sufficient to compute value-added scores and produce student growth reports for 80 percent of the teachers, who were in grades 4-8.
- **Although most teachers and principals participated in the student growth report training, less than half of the teachers and principals accessed their reports.** Overall, 85 percent of teachers and 81 percent of principals participated in a webinar prior to the release of the student growth reports. The webinar oriented the participants to the value-added scores, the content of the student growth reports, and how to access them. The online reporting system showed that 40 percent of the teachers with value-added scores and 38 percent of the principals accessed their student growth reports.

What Were the Characteristics of the Student Growth Performance Information Provided?

- **Student growth reports included school and individual teachers' value-added scores.** The teacher report included a teacher's overall and subject-specific value-added scores (both reading/English language arts and mathematics for those who taught both subjects) with an indication of their percentile ranking relative to other teachers in the district, and the average teacher score in the district and school. All scores included confidence intervals/standard error information to indicate the precision of the estimated scores. Each teacher could also access a roster that included the number and names of students used to calculate their score. For each teacher in his or her school, the principal could view an overall value-added score, scores by subject and grade, and scores across time. Principals could also view school average scores overall and by subject and grade.
- **Many teachers with a student growth report had a value-added score that measurably differed from the district average, particularly in mathematics.** The student growth reports available to teachers and principals included teachers' 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 teacher.[16] For example, in mathematics, 25 percent of the teachers had a value-added

[15] 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.

[16] 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 is actually

score that was considered measurably below the district average, and 28 percent had a score that was considered measurably above average. See exhibit ES.3.

Exhibit ES.3. Distribution of treatment teachers based on whether their value-added score was considered measurably above or below the district average, by subject

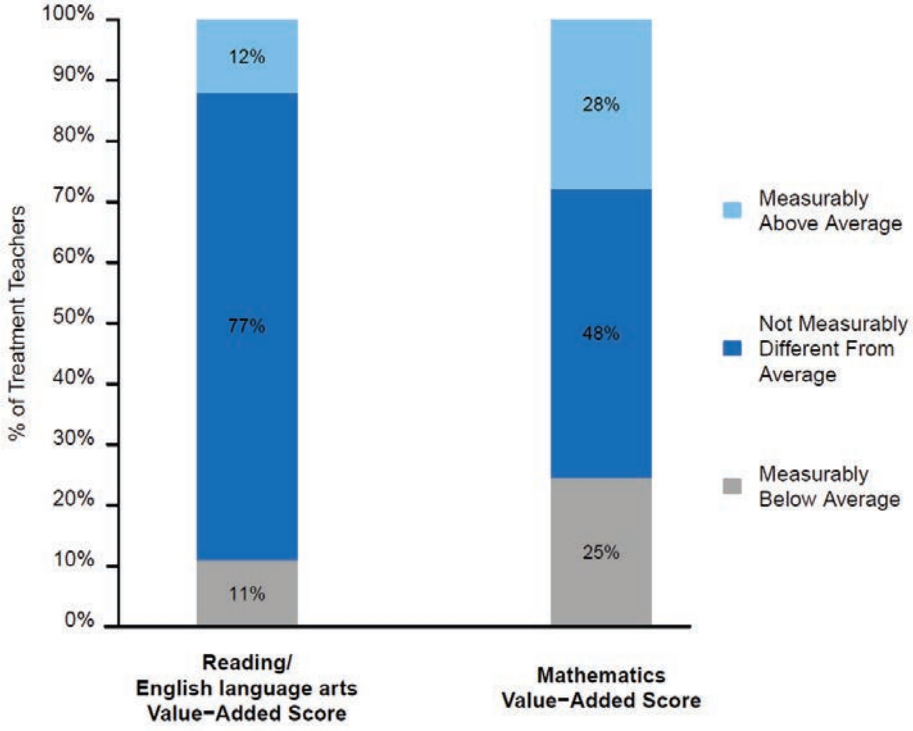


Exhibit Reads: For treatment teachers with mathematics value-added scores, 28 percent had scores considered measurably above the district average.

NOTE: Distributions of teachers are based on whether the 80 percent confidence interval for a teacher's value-added score was above or below the district average. To indicate the amount of uncertainty around each teacher's score, the student growth reports included 80 percent confidence intervals, which showed the range of scores that have an 80 percent chance of including the teacher's "true" score. This benchmark was selected in order to appropriately balance two types of risks within the context of an intervention designed to provide feedback on performance without explicit consequences such as promotion or dismissal: (1) the risk of misidentifying truly average teachers as below- or above-average, and (2) the risk of misidentifying teachers who were truly below- or above-average as average teachers. Sample size = 338 teachers with mathematics value-added scores and 321 teachers with reading/English language arts value-added scores. Reported percentages may not sum to 100 percent because of rounding.

SOURCE: AIR value-added system.

The Principal Leadership Measure and Feedback

Feedback on principal leadership 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. The VAL-ED includes six "core components" of principal performance: high standards for student learning, rigorous curriculum, quality instruction,

average as above or below average, against the risk of misclassifying a teacher who is 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.

culture of learning and professional behavior, connections to external communities and performance accountability. Principals are also rated on six "key processes": planning, implementing, supporting, advocating, communicating, and monitoring. A report for each principal was generated after each administration of the VAL-ED, and the principal's supervisor was expected to discuss the report with the principal in a feedback session.

How Was the Principal Leadership Measure Implemented?

- **All principals and their supervisors received training on using VAL-ED.** All principals and their supervisors participated in a two-hour VAL-ED training in summer 2012. During the school year, all principals' supervisors also received a one-hour training to prepare them to conduct the feedback sessions. In addition, teachers were offered a one-hour introduction to VAL-ED at the beginning of the school year, as well as an orientation webinar during the school year.
- **All VAL-ED reports incorporated input from the principal, the principal's supervisor, and most teachers.** All principals and their supervisors completed the VAL-ED rating form, and a high percentage of teachers in each treatment school (80 percent in fall and 90 percent in spring on average) also completed the form.
- **All VAL-ED feedback sessions occurred as planned.** In both fall and spring, all principals met with their supervisors to discuss their VAL-ED reports. Principal supervisors reported feedback sessions lasting on average 52 minutes in the fall and 46 minutes in the spring.

What Were the Characteristics of the Principal Leadership Performance Information Provided?

- **The VAL-ED reports present scores and performance levels, as well as percentile ranks, for each dimension of leadership.** VAL-ED reports present an overall score, a score for each core component, and a score for each key process. For each of these 13 scores, the report additionally presents a performance label and a percentile rank, relative to the principals included in a national VAL-ED field test. Each score (i.e., overall score, core component scores, and key process scores) is an average across the three respondent groups (i.e., principal, supervisor, and teachers), with each group weighted equally. The report additionally shows the scores received from each respondent group separately.
- **The VAL-ED ratings classified some principals as lower-performing and some as higher-performing.** In the fall, principals' overall scores were distributed across the four performance levels (8 percent of principals were labeled *distinguished*, 22 percent *proficient*, 43 percent *basic*, and 27 percent *below basic*). In the spring administration, half the principals received an overall score associated with a performance level of *proficient* or *distinguished* and half received a score at the *basic* or *below basic* level.[17] (See exhibit ES.4.)

[17] The increase in average VAL-ED overall scores from the fall to spring is primarily a product of an increase in the principal self-ratings. Average ratings of principal leadership based on the three respondent groups were similar in

Exhibit ES.4. Distribution of treatment principals across performance levels based on VAL-ED overall scores, by assessment window

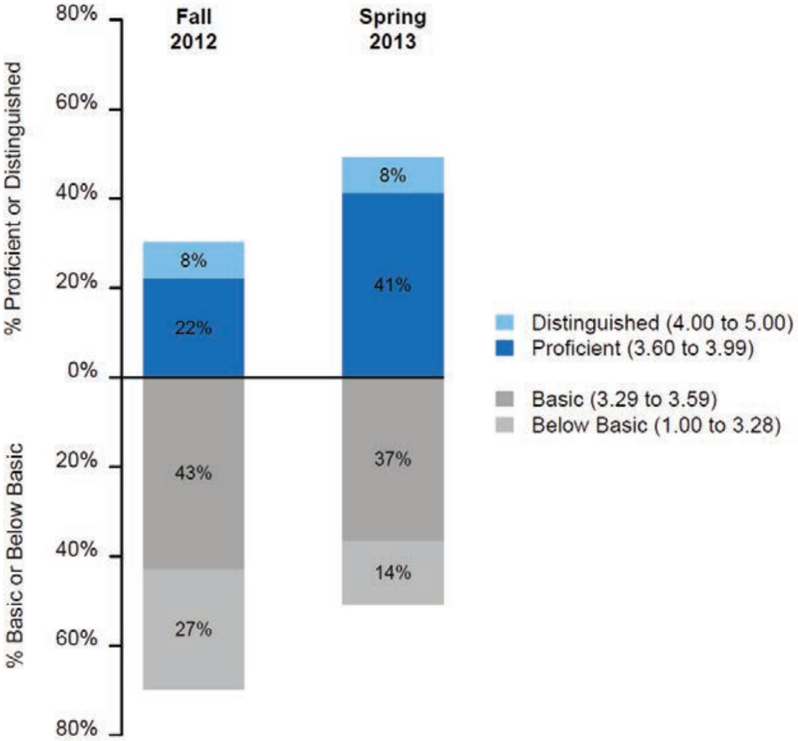


Exhibit Reads: In fall 2012, 8 percent of treatment principals had a VAL-ED overall score at the *distinguished* performance level, 22 percent at the *proficient* level, 43 percent at the *basic* level, and 27 percent at the *below basic* level.

NOTE: Performance level distributions are based on principals' VAL-ED overall scores at each assessment window. The overall score is an average of the scores from the principal's supervisor, teachers, and the principal's own self-rated score, with each group weighted equally. Sample size = 63 principals for both fall 2012 and spring 2013. Reported percentages may not sum to 100 percent because of rounding.

SOURCE: Fall 2012 and Spring 2013 VAL-ED Surveys.

- **VAL-ED ratings provided by principals, supervisors, and teachers in the fall were often too different to form a reliable measure, but the spring ratings were consistent enough to distinguish between some lower- and higher-performing principals.** To provide information about a principal's overall effectiveness, the VAL-ED scores should communicate a consistent (i.e., reliable) message about the principal's effectiveness across the three respondent groups (the principal, the principal's supervisor, the principal's teachers). Based on the literature on 360-degree surveys, we would expect correlations between respondent group scores between .25 and .35.[18] In the fall, however, agreement among the three respondent groups' overall scores was low, with correlations ranging from .06 to .27. In the spring, correlations were higher (between .26 and .38), and thus the reports provided a more consistent message

the fall; however, in the spring, principal self-ratings were higher on average (3.76) than the ratings from their supervisors (3.50, p-value of the difference <.05) and teachers (3.57, p-value of the difference < .05).

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

about a principal's effectiveness. Viewing and discussing the fall reports may have led principals and their supervisors to better align their ratings in the spring.

Educators' Performance Evaluation Experiences

The study's performance measures were intended to provide educators with performance information that was more frequent, systematic, and useful as a guide for professional growth than the information that they normally receive. To assess whether this occurred, we compared the treatment and control groups' responses on surveys administered in the spring. Teacher surveys were usually completed at the beginning of the last of the four observation windows. Principal surveys were completed prior to the spring VAL-ED feedback session, which usually occurred at the end of the school year.

What Were Teachers' Experiences?

- **Treatment teachers reported receiving more feedback on both their classroom practice and their students' achievement growth than control teachers.** Treatment teachers reported receiving more feedback sessions with ratings and a written narrative than control teachers (3.0 versus 0.7 instances). The average treatment teacher also received a larger amount of oral feedback than the average control teacher (80 minutes versus 18 minutes). Furthermore, relative to control teachers, treatment teachers were more likely to report receiving value-added scores (45 percent versus 24 percent) and less likely to report receiving test scores for individual students or classroom average scores.[19]
- **Among those who reported receiving feedback, treatment teachers indicated somewhat more positive perceptions than control teachers about the information they received on their classroom practice but not about the information on their students' achievement.** Although most teachers in both treatment and control groups reported agreeing or strongly agreeing that the feedback on their classroom practice provided specific ideas about how to improve, treatment teachers were more likely to report so (87 percent versus 79 percent). Almost all teachers (approximately 92 percent) in both groups indicated that the feedback on classroom practice was a fair assessment of their performance. Control teachers were more likely than treatment teachers to report that the student achievement information they received was easy to understand (89 percent versus 78 percent). However, less than half of the teachers in both groups agreed or strongly agreed that the achievement information was a fair assessment of their performance (49 percent for treatment teachers and 43 percent for control teachers, not a statistically significant difference) or a fair indicator of teacher effectiveness for all teachers (40 percent for treatment teachers versus 29 percent for control teachers, a statistically significant difference).

[19] This finding should be interpreted with caution because some teachers may not have had a correct understanding of the term "value-added scores." As a validity check, we compared treatment teachers' responses with electronic records indicating which teachers had accessed their own value-added scores in the online system, and we found that 34 percent of the treatment teachers who reported receiving value-added scores did not access their student growth reports in the online system, and 17 percent of treatment teachers who reported not receiving value-added scores actually accessed their online student growth reports.

What Were Principals' Experiences?

- **Treatment principals reported receiving more feedback than control principals.** Treatment principals reported receiving feedback more often than control principals (2.0 versus 1.4 instances) and more instances of oral feedback with ratings (1.0 versus 0.4 instances). The average treatment principal also received more oral feedback than the average control principal (60 minutes versus 41 minutes). However, treatment principals were no more likely than control principals to report that their supervisors' feedback focused on specific topics related to VAL-ED.
- **Among those who reported receiving feedback, most principals in both treatment and control schools had positive perceptions about the feedback they received.** The majority (more than 70 percent) of the principals in both treatment and control schools agreed that the feedback they received was a fair assessment of their performance, and approximately two thirds or more of the principals agreed that the feedback they received contained specific ideas for improving their performance. Among those who received feedback, there was no statistically significant difference between treatment and control principals in their perceptions of the feedback.

Future Report

This report focuses on findings from the first year of implementation of the study's three performance measures with feedback. Findings about the second year of implementation will be presented in the second-year study report. The second-year report also will present findings on the impact of the study's performance measures and feedback on teacher classroom practice, principal leadership, and student achievement.

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

Chapter 1. Introduction

Educator performance evaluation systems are a potential tool for improving student achievement.[20] By removing ineffective teachers and principals and/or through increasing the effectiveness of the existing workforce, such systems may result in higher student achievement.[21]

Emerging research suggests some promising features of performance evaluation measures. For example, research suggests that, to measure classroom practice, additional observations of the same teacher, beyond the first, form a more reliable measure of a teacher's typical practice, especially when more than one observer is used.[22] There is also some evidence from recent research that giving more frequent, specific feedback on classroom practice may lead to improvements in teacher performance and student achievement.[23]

The U.S. Department of Education's Institute of Education Sciences is conducting a study on the implementation and impacts of teacher and principal performance measures that are highlighted by emerging research. 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

The study has two main goals. The first is to examine the implementation of the intervention in a set of districts, including how well it was implemented and the characteristics of the performance measures. The second goal is to examine whether the intervention affected educator outcomes (e.g., teachers' classroom practice) and, ultimately, student achievement, when implemented in districts with evaluation system practices that are less objective and intensive than the intervention.

This report focuses on the first year of the two years of implementation, describing the characteristics of the educator performance measures and teachers' and principals' experiences

[20] See Stecher et al. (2016); Weisburg, Daniel, Sexton, Mulhern, and Keeling (2009).

[21] Researchers studying a range of educator workforce interventions theorize that there are two key mechanisms leading to improved student achievement: changes in the composition of the workforce and changes in the skills of continuing workers. See, for example, Taylor and Tyler (2012) regarding teacher evaluation, Chiang et al. (2015) regarding performance-based compensation, and Glazerman et al. (2010) regarding comprehensive teacher induction.

[22] See Bill & Melinda Gates Foundation (2012); Whitehurst, Chingos, and Lindquist (2014).

[23] See Steinberg and Sartain (In Press); Taylor and Tyler (2012).

with performance feedback. It is one of the few detailed descriptions of educator performance measures implemented on a large scale in districts.[24] The final report will primarily focus on impacts on outcomes including principal leadership, teacher classroom practice, and student achievement but will also include descriptive information on the second year of implementation.

This chapter describes the study's intervention, research questions, and design.

Overview of the Intervention

The intervention consisted of three performance measures that were implemented in tandem, providing feedback to those being evaluated and their supervisors. The intervention was intended to have many of the features promoted by research, specifically:

- Multiple measures of teacher and principal performance, including classroom observations and student growth.
- Measures that provide meaningful information about differences in educator performance (i.e., the measures vary across individuals and are reliable).
- Measures that provide feedback that is clear and useful at multiple times during the year.[25]

In each of the eight participating districts, the intervention was implemented in select elementary and middle schools. A group of control schools in each district participated in the normal evaluation processes only.

The intervention specified how educators would receive the feedback (e.g., in feedback sessions after each observation). Other potential uses of the performance information were left to the discretion of the participating school and central office staff. The study's implementation team held meetings in each district to ask a group of school and central office educators to consider ways the performance information might be used, such as to identify educators for praise or support, to plan professional development, or to guide coaching. Although districts were given the option of using the information for staffing decisions such as tenure or continued employment, the study team anticipated that these uses might be difficult, since using feedback for high stakes purposes might require changes to contracts or other agreements that could not be made quickly. The districts decided not to use the information in this way, for the most part; in three districts, the observations conducted by principals as part of this study counted in their official rating system if the teacher was due to be observed that year under the district's existing evaluation system.

Thus, the study tests the impact of providing feedback without stakes attached, as an add-on to existing performance feedback. The available research evidence is mixed on whether stakes increase the effectiveness of feedback or attenuate it. Some analysts hypothesize that employees may be more motivated to change their practices if they view their evaluation system as being

[24] For an example, see Lipscomb, Terziev, and Chaplin (2015).

[25] Bill & Melinda Gates Foundation (2012); Bill & Melinda Gates Foundation (2013); Whitehurst, Chingos, and Lindquist (2014).

used for purposes of professional development, instead of dismissal (e.g., Smither et al., 2005, and Atwater et al., 2007). On the other hand, two recent studies in districts that provide feedback similar to that provided by this study's intervention found that attaching stakes to the feedback had a positive effect.[26]

Below we describe each of the three intervention performance measures.

1. The Teacher Classroom Practice Measure and Feedback

This performance measure used classroom observations four times during the course of the year, with a feedback session after each observation.[27] One of the four observations was intended to be conducted by an administrator from the teacher's building, and the other three were intended to be conducted by study-hired observers (i.e., local professionals hired and trained by the study).[28]

After each observation, the observer was expected to prepare a report including both ratings and narrative feedback on the teacher's classroom practice. The observer was also expected to hold an in-person feedback session, within one to two weeks, lasting approximately 45 minutes, to review the report with the teacher. To ensure that building administrators received the performance information for all teachers in their respective schools, the study held midyear meetings of the building administrators in each district in the winter to review classroom practice reports and learn how to access future reports through a secure online portal.

Two different classroom observation systems were used. Districts were asked to choose between the Classroom Assessment Scoring System (CLASS) and Charlotte Danielson's Framework for Teaching (FFT). The treatment schools in four of the eight study districts used the CLASS, and the treatment schools in the other four study districts used FFT.[29] The use of two different observation systems was intended to make the study findings more broadly relevant than would

[26] Chiang et al. (2015) found that attaching compensation to the evaluation system performance measures had an impact on student achievement in reading but not mathematics. Dee and Wycoff (2013) examined the impact of attaching the threat of dismissal for low-performance, and, separately, of attaching a prospect of a large financial bonus for sustained high-performance. Using a regression discontinuity design, it found that both affected teachers' performance ratings. These studies were done in districts that provide feedback to all teachers similar to that provided by this study's intervention and focused on the stakes attached to that feedback.

[27] In addition to four observations per year for the teachers who were the focus of the study (i.e., grades 4–8 teachers responsible for mathematics and reading/ELA instruction), the performance measure was used to provide two observations per year for K–3 teachers—one by the principal and one by a study-hired observer. These additional observations were intended to foster a sense of collective participation in the implementation of the classroom practice performance measure in the participating elementary schools, as there is some evidence suggesting that collective participation in professional development initiatives may enhance their chances for success (see Garet et al. 2001). In the middle schools, no additional observations were done, as departmentalized teachers may already have a sense of collective participation through the participation of others in their department. The appendixes contain supplemental tables with results for K–3 teachers.

[28] This distribution of effort was intended to engage principals in the implementation of the performance measure without overburdening them. In addition, compared with using a single observer for each teacher, the use of multiple observers to rate the same teacher produces a more reliable end-of-year average (see Ho and Kane 2013).

[29] Several districts recruited for the study indicated that they were indifferent between CLASS and FFT, and they were assigned as needed to achieve the intended balance.

be the case if only one system were used. However, the districts were not randomly assigned to the two systems, so the study design does not allow us to draw conclusions about their relative effectiveness.

CLASS and FFT share many features that made them suitable for the study. First, they focus on similar dimensions of instruction, and the rating levels on each dimension are defined using specific, observable behaviors of teachers and students. In addition, for both instruments, there is evidence of validity and an association with student achievement (Allen et al. 2011; Bill & Melinda Gates Foundation 2012; Goe, Bell, and Little 2008; Mashburn et al. 2010). Both instruments are also applicable across subjects and grades.

CLASS and FFT were also suitable for the study because support for implementation of CLASS and FFT was available from national vendors. The study contracted with these vendors, who provided the standard observer training to the observers (i.e., the principals and study-hired observers). Each trained observer had to demonstrate sufficient skill in rating on a video-based assessment. The vendors also provided related trainings, materials, and Web-based platforms for managing and reporting the performance information.[30] Among the materials made available to teachers were online video libraries that teachers and observers could access to view examples of teaching that exemplify particular levels of performance on each measured dimension of classroom practice.

2. The Student Growth Performance Measure

This performance measure was based on student test results from multiple years to provide information about each teacher's contribution or the "value added" to student academic growth. A value-added score is an estimate, based on a statistical model, of how a teacher's students performed during the year, on average, compared with similar students in the district (i.e., those in the same grade with similar prior performance and other characteristics). Teacher value-added scores are positively related to teacher instructional practices (Grossman et al. 2013; Hill, Kapitula, and Umland 2011). In addition, there is some evidence that a teacher's value-added score is a valid predictor of student academic achievement (Chetty, Friedman, and Rockoff 2014a; Kane, McCaffrey, Miller, and Staiger 2013; Kane and Staiger 2008) and longer-term student outcomes (Chetty, Friedman, and Rockoff 2014b).

During the two years of the study, AIR prepared three waves of value-added reports, each focusing on a different period of instruction. The first wave of reports was released between February and April of the first study year. The second and third waves were released in the fall of the second study year and the fall of the year after the study.

Because computing value-added scores requires that students have at least one pretest score, the student growth performance measure focused on teachers of grades 4-8 who were responsible for instruction in mathematics and reading/English language arts (ELA). All of the study districts had sufficient data to compute value-added scores in these grades.

[30] The organizations who provided support for the CLASS version of the classroom practice performance measure were Teachstone and the University of Virginia. The organizations who provided support for the FFT version of the classroom practice performance measure were Danielson Group and Teachscape.

An AIR team separate from the evaluation team designed and conducted the value-added analysis, based on AIR's experience doing similar work for states and input from members of the study's technical working group. Value-added scores were generated for each teacher using a covariate adjustment model, an approach widely used in current state and district implementations of value-added (see Collins and Amrein-Beardsley 2014). The model used for each district incorporated student test scores for two prior years (where available) as predictors, along with a set of measures of student characteristics selected by the districts. This choice of model and other design decisions were based on three design criteria: (1) the statistical model should produce technically defensible scores, (2) the approach should minimize data requirements to include as many teachers and their students as possible while maintaining its technical rigor, and (3) the approach should allow some district-specific adjustments to align with district context and policy. (See Appendix F for technical details about the estimation of value-added scores for the intervention.)

3. The Principal Leadership Performance Measure and Feedback

The principal leadership performance measure was designed to provide principals and principal supervisors with feedback on principal leadership, which was measured twice a year (fall and spring) using the Vanderbilt Assessment of Leadership in Education (VAL-ED). VAL-ED is a 360-degree survey that assesses principal leadership from the perspectives of the principal, the principal's supervisor, and teachers. It was selected for this study because it is aligned with national standards for principal leadership (Goldring et al. 2009) and because it has demonstrated validity and reliability (Condon and Clifford 2010).[31] After each survey administration, the VAL-ED vendor, Discovery Education, generated a report on each principal with detailed survey results. The principal and the principal's supervisor were then to hold a one-on-one feedback session to discuss the results.

To prepare them to implement all three of the performance measures, teachers, principals, and principal supervisors received trainings from the vendors, as described in the chapters that explain each measure in detail. In addition, teachers received a one-day orientation just prior to the beginning of the first study year. The orientation day included three hours on the intervention's measure of classroom practice to help teachers begin to understand the focus of the classroom observation instruments. The orientation day also included one hour on the measure of student growth and one hour on the measure of principal leadership.

Theory of Action and Research Questions

This study is guided by a theory of action that is based on hypotheses about how performance measures and feedback may affect the outcomes of teachers, principals, and students (see exhibit 1.1). According to the theory, frequent and systematic performance measurement and feedback may generate information that distinguishes between lower- and higher-performing educators and between different dimensions of an individual educator's performance, which could help identify educators in need of support and dimensions on which an educator should improve.

[31] The researchers who developed VAL-ED have published its psychometric properties in peer-reviewed journals and their website (http://www.valed.com/research.html). See, for example, Porter et al. (2010).

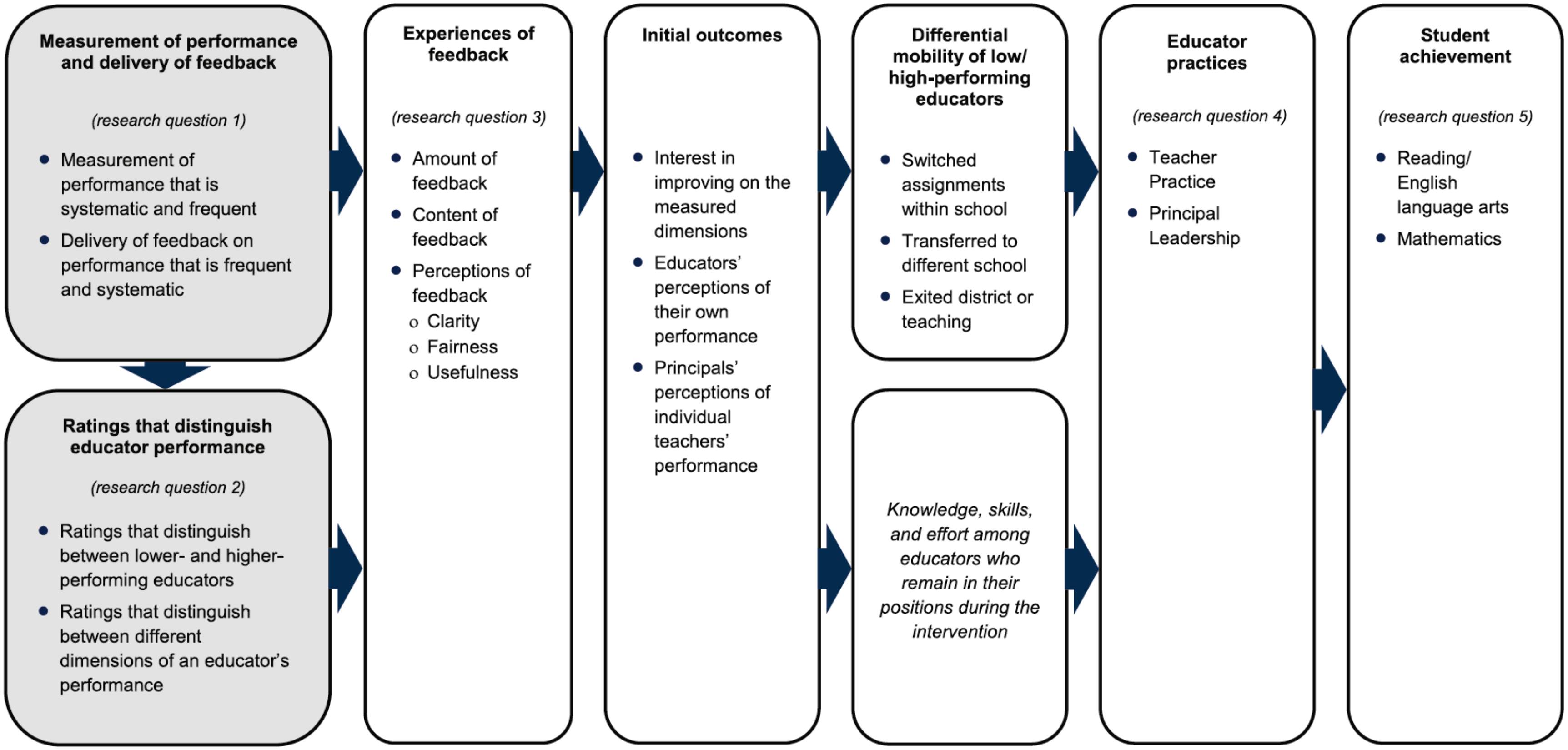
We hypothesize that if the feedback is frequent and perceived as clear, fair, and useful, it may have an impact on educators' "initial outcomes," most immediately their interest in improving along the dimensions on which they received feedback. This may lead teachers and principals to get support for improvement, for example through professional development, consulting colleagues, or independently identifying and implementing new classroom practices. [32]

These effects on educators may, in turn, affect teacher classroom practice and principal leadership through two distinct mechanisms. First, a change in perceptions about performance, in either self-perceptions or the supervisor's perceptions, could result in differential mobility between low- and high-performing educators, changing the overall composition of the educator workforce. Second, performance measurement and feedback could have an impact on educator practices and student achievement by leading to improved knowledge, skills, and effort of teachers and principals who remain in their positions during the intervention. Thus, through either or both of the two mechanisms, performance measurement and feedback could have a positive impact on the quality of teacher classroom practice and principal leadership, and, in turn, on student achievement, as shown in the far right of the theory of action diagram.[33]

[32] There is some evidence that feedback can lead to improvements in classroom practice. For example, a professional development program designed to provide frequent feedback based on the secondary school version of the CLASS rating instrument had an effect on student achievement that was partially mediated by classroom practice (Allen et al. 2011). There is little evidence, however, on the intermediate mechanisms that lead to improved classroom practice, or on the features of feedback needed to lead to improvements in classroom practice. Title II of the Every Student Succeeds Act allows the use of federal funds to support the design and implementation of performance evaluation systems that provide feedback that is "clear, timely, and useful."

[33] For literature discussing these mechanisms, see footnote 21.

Exhibit 1.1. Theory of action



Measurement of performance and delivery of feedback

(research question 1)

- Measurement of performance that is systematic and frequent
- Delivery of feedback on performance that is frequent and systematic

Ratings that distinguish educator performance

(research question 2)

- Ratings that distinguish between lower- and higher-performing educators
- Ratings that distinguish between different dimensions of an educator's performance

Experiences of feedback

(research question 3)

- Amount of feedback
- Content of feedback
- Perceptions of feedback
  - o Clarity
  - o Fairness
  - o Usefulness

Initial outcomes

- Interest in improving on the measured dimensions
- Educators' perceptions of their own performance
- Principals' perceptions of individual teachers' performance

Differential mobility of low/ high-performing educators

- Switched assignments within school
- Transferred to different school
- Exited district or teaching

Knowledge, skills, and effort among educators who remain in their positions during the intervention

Educator practices

(research question 4)

- Teacher Practice
- Principal Leadership

Student achievement

(research question 5)

- Reading/ English language arts
- Mathematics

This multiyear study is designed to examine the implementation of an intervention that is guided by this theory of action and estimate its impact on educator and student outcomes. It addresses five research questions:

1. To what extent were the performance measures and feedback implemented as planned?
2. To what extent did the performance measures distinguish educator performance?
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?

This report will address the first three questions, focusing on the first year of implementation.

Overview of Study Design

To answer the study's research questions, we recruited a sample of eight districts and conducted the study in a selected group of schools in each district. The participating schools were assigned by lottery to implement the study's intervention (the treatment group) or not (the control group). The treatment group implemented the study's intervention, although both continued to implement the districts' existing educator evaluation systems. In the participating schools, the study focused on the principals and teachers of mathematics and reading/English language arts in grades 4-8.[34]

This section describes the sample and how it was selected, how we randomly assigned schools to treatment and control, what data we collected, and what analytic methods we used.

Sample Selection

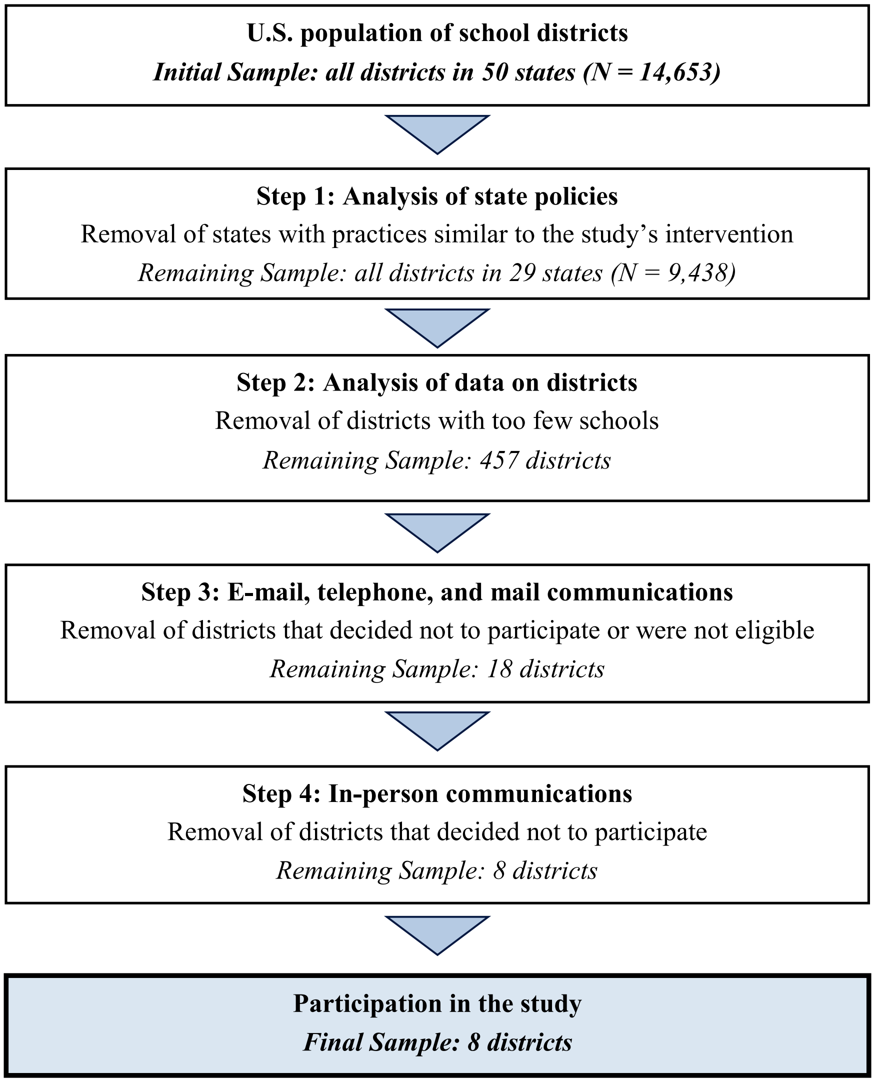
The district selection process took place between October 2011 and May 2012, and it resulted in a final study sample of eight districts where existing policies for the evaluation of teachers and principals contrasted with the study's intervention. The process began, as shown in exhibit 1.2, with an analysis of state policies for the evaluation of teachers and principals. Several states (e.g., many of the states with Race to The Top grants) had begun to implement practices that were similar to the study's intervention or planned to implement such practices before the end of the study's two-year implementation period, from fall 2012 to spring 2014. The study team excluded districts from those states. Because of the ESEA Flexibility Waivers, many other states planned that districts would implement such practices, but not until fall 2014. Thus, districts in many states were eligible despite the state's participation in the waiver program).

[34] 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 are not included in the main study analyses, however, because student assessment data needed for the feedback on student growth (i.e., needed to calculate value-added scores) are not available in Kindergarten through grade 3. In addition, the assessment data required to analyze the impact of the intervention on student achievement are not available in Kindergarten through grade 2.

Within the remaining 29 states, there were 457 districts that met the study size criteria of at least 20 elementary and middle schools, based on information from the 2009-10 *Common Core of Data.* Attempted e-mail, telephone, and mail communications with the 457 districts led to initial conversations with 100 districts, and 49 expressed interest in a follow-up conversation about participating in the study. The study team assessed district eligibility and determined that some were not eligible (i.e., they did not have data systems that made the student growth performance measure feasible, or they had policies for evaluation of teachers and principals that did not contrast with each of the intervention's three performance measures). Of the 36 that were eligible, 18 were interested in an in-person meeting.

AIR visited all 18 remaining districts and held a recruitment conference in Washington, D.C., for districts that continued to be interested in participation. Thirteen districts were sufficiently interested to attend the recruitment conference. Of these, five eventually declined participation, for a combination of reasons that differed by district (such as likely objection by the teacher's organization or the aggressive schedule to begin implementing the intervention performance measures in summer 2012).

Exhibit 1.2. District selection and recruitment process



U.S. population of school districts

Initial Sample: all districts in 50 states (N = 14,653)

Step 1: Analysis of state policies

Removal of states with practices similar to the study's intervention

Remaining Sample: all districts in 29 states (N = 9,438)

Step 2: Analysis of data on districts

Removal of districts with too few schools

Remaining Sample: 457 districts

Step 3: E-mail, telephone, and mail communications

Removal of districts that decided not to participate or were not eligible

Remaining Sample: 18 districts

Step 4: In-person communications

Removal of districts that decided not to participate

Remaining Sample: 8 districts

Participation in the study

Final Sample: 8 districts

Characteristics of the Study Districts

At the conclusion of the recruitment process, the sample included eight districts that spanned all geographic regions except the Northeast, with two or three districts in each region (see the right-hand column of exhibit 1.3). Many states in the Northeast were deemed ineligible because they had accepted federal or foundation grants to reform their evaluation systems during or before the study's implementation period.

The sample was also decidedly urban (75 percent versus 7 percent nationally), including only one suburban and one rural district, mostly because of the removal of districts that did not have the required number of schools.

Exhibit 1.3. Characteristics of all districts in the United States and districts that participated in the study

District characteristics	All districts in the United States	Districts that participated in the study
Geographic region (percentage of districts)		
Midwest	36.1	37.5
Northeast	21.0	0.0
South	23.0	37.5
West	20.0	25.0
Urbanicity (percentage of districts)		
Urban	6.7	75.0
Suburban	19.9	12.5
Town	17.3	0.0
Rural	56.1	12.5
Number of schools	6.5	39.3
Number of full-time equivalent teachers	202.7	1,255.7
Total enrollment	3,470.3	19,995.4
Title I eligible (district average percent of schools)	72.3	58.5
Free or reduced-price lunch (district average percent of students)	34.1	31.2
Race/ethnicity (district average percent of students)		
Asian	2.0	2.6
African American	7.3	3.5
Hispanic	13.0	41.4
White	72.4	48.4
Other	5.3	4.2
State requires collective bargaining (percentage of districts)	67.7	37.5
Number of Districts	**14,653**	**8**

NOTE: Percentage values for characteristics with multiple categories may not sum to 100 because of rounding.

SOURCE: 2011–12 Common Core of Data; National Council on Teacher Quality Teacher Contract Database (retrieved in May 2015).

The sample spanned a range of state policies with respect to collective bargaining: Three districts were in states where collective bargaining is illegal, and the other districts were in states where collective bargaining was permissible or required. By comparison, across the United States 68 percent of districts are in states where collective bargaining is required. The loss of districts in states requiring collective bargaining occurred during the final step of the recruitment process. Although it is not possible to know districts' reasons for dropping out, it was common for districts with collective bargaining agreements to consider teacher union support as a factor in the decision.

Performance Feedback Typically Provided in the Study Districts

By design, the performance feedback provided as part of the study's intervention was to be in addition to the feedback typically provided by districts. We conducted interviews with each district to determine what the districts typically provided. (The interviews are described further below under "Data Collection" and in appendix B.) The districts' feedback to teachers and principals on classroom practice, student growth and principal leadership differed from the feedback planned as part of the study's intervention.

Districts' feedback on classroom practice. In all eight study districts, the districts required less frequent observation of teachers than the intervention's four observations per year. Most districts required observations of nonprobationary teachers, the majority of the teacher sample, less frequently than once a year. Across the study districts, requirements for observations of nonprobationary teachers ranged from once a year to once every five years, averaging about once every two years. (See Exhibit 1.4.)

District policies also differed from the study intervention in who conducted observations. Under the districts' evaluation systems, building administrators conducted the observations. By contrast, the intervention used study-hired observers for three of the four observations each year.

In addition, district policies differed from the intervention in the training requirements for observers. The districts required an average of 13.5 hours of training, or a little over half of the duration of the study's training.[35] In two of the eight districts, no observer training was required. Only three districts required observers to pass an assessment of rating skill, which is required for the study's intervention.

District policies were somewhat similar to the intervention in one respect: Each of the study districts used a classroom observation instrument that, like the study's observation instruments (CLASS and FFT), measured classroom practice on several dimensions and defined multiple performance levels for each dimension. In five of the districts, the instrument was an adaptation of the FFT; these districts, for instance, changed the names of the performance levels, or altered the text that defines the performance levels for each dimension.

Districts' feedback on student growth. In contrast to the intervention, none of the districts provided value-added scores to teachers, nor did their state education agencies (see Exhibit 1.4).

[35] The required observer training for the study's intervention was 20 hours for observers in CLASS districts and 26 hours in FFT districts.

Exhibit 1.4. Policies and practices for performance feedback to teachers, by district

Districts' Feedback on Teachers' Classroom Practice								Districts' Feedback on Student Growth	
District number and assigned classroom observation system for intervention		**Frequency of observation with feedback[a]**		**Use of staff not based at the school as observers**	**Features of observer training**		**Use of rating instrument that differentiates at least 3 performance levels and provides ratings for multiple dimensions of performance**	**Value-added scores provided to teachers**	**Information on changes in achievement provided to teachers[c]**
		Probationary teachers[b]	*Nonprobationary teachers[b]*		*Duration of required training*	*Required assessment of rating skill*			
1	CLASS	1 per year	1 every three years	No	9 hours	No	Yes, adapted FFT	No	No
2	CLASS	1 per year	1 every five years	No	40 hours	Yes	Yes	No	Yes
3	CLASS	1 per year	1 every two years	No	24 hours	Yes	Yes	No	Yes
4	CLASS	3 per year	1 per year	No	None	No	Yes, adapted FFT	No	Yes
5	FFT	2 per year	1 every three years	No	4 hours	No	Yes, adapted FFT	No	Yes
6	FFT	2 per year	1 every two years	No	7 hours	No	Yes, adapted FFT	No	Yes
7	FFT	2 per year	1 per year	No	None	No	Yes, adapted FFT	No	Missing
8	FFT	1 per year	1 every four years	No	24 hours	Yes	Yes	No	Yes
Overall average		1.6 per year	0.5 per year		13.5 hours				

[a] Number of observations shown is the minimum required under each district's evaluation system. Administrators could observe more frequently at their discretion.

[b] Each of the eight study districts categorized teachers as probationary or nonprobationary in part on the basis of service in the district. In most of the districts, probationary teachers were eligible to become nonprobationary after three years of service; in the other districts, they were eligible after one year of service. Across the sample, 15 percent of grades 4–8 teachers had three or fewer years of experience as teachers in their district.

[c] The six districts indicated that this information was provided to teachers routinely for informational purposes rather than performance measurement. One district reported that such information was not provided, and one district did not respond.

SOURCE: District interview

Although six districts provided teachers with information on changes in their students' achievement to monitor individual student progress (e.g., changes during the year based on quarterly diagnostic tests), this did not include information that would necessarily provide teachers with a sense of their teaching performance.

Districts' feedback on principal leadership. In all eight study districts, feedback on principal performance was required once a year, in contrast to the study intervention's plan of twice a year. (See Exhibit 1.5.) District policies for principal evaluation also differed from the intervention in the nature of the information used for feedback. None of the districts used the VAL-ED instrument, which was the study's principal performance measure. And, only two districts systematically collected teacher input on principal performance through a survey, which is a key feature of the VAL-ED. Finally, district policies were similar to the intervention in one respect: each of the study districts measured principal performance on multiple dimensions, and at least six of the districts rated principals on three or more performance levels.

Exhibit 1.5. Policies and practices for performance feedback to principals, by district

		Districts' Feedback on Principal Leadership			
District number and assigned classroom observation system for intervention		Frequency	Use of teacher survey as input in principal evaluation	Rating instrument with multiple dimensions	Performance on each dimension rated using three or more performance levels[a]
1	CLASS	1 per year	No	Yes	Yes
2	CLASS	1 per year	No	Yes	Yes
3	CLASS	1 per year	No	Yes	Missing
4	CLASS	1 per year	Yes	Yes	Missing
5	FFT	1 per year	No	Yes	Yes
6	FFT	1 per year	No	Yes	Yes
7	FFT	1 per year	Yes	Yes	Yes
8	FFT	1 per year	No	Yes	Yes

[a]Data for two districts are missing because the districts did not provide the rating instruments.

SOURCE: District interview.

School Selection and Characteristics of the Study Schools

Each of the eight districts identified a set of schools that met the study's eligibility criteria and agreed to participate in the study. Because of the study's focus on teachers of reading/ELA and mathematics in grades 4–8, schools eligible for the study were elementary and middle schools. To reduce heterogeneity in the school sample, the sample was restricted to regular schools, operated by the school district (i.e., noncharter schools).

Consistent with the characteristics of the study districts, the participating schools were similar to schools in the national population in terms of enrollment and Title I status, but they differed in other characteristics such as urbanicity and student demographic composition. Compared with the national population, for example, schools in the study sample were more likely to be urban and had a higher percentage of students who were minorities on average. (See exhibits A.1 and A.2 in appendix A. For the characteristics of schools in the districts that used CLASS and FFT, see exhibit A.3 in appendix A.)

Random Assignment

The participating schools were assigned by lottery to implement the study's intervention (the treatment group) or not (the control group). Both groups continued to implement their district's existing educator evaluation systems, although the treatment group also implemented the study's intervention.

To maximize the precision with which the study could compare outcomes in the treatment and control group, random assignment was conducted separately within 37 blocks. The blocks were defined by district and school level (elementary schools or middle schools). Thus, half of each district's elementary schools were treatment schools and half were control; likewise, half of each district's middle schools were treatment and schools and half were control. Blocks additionally took into account school size and/or the percentage of students eligible to receive free or reduced-price lunch.

In total, 63 treatment schools and 64 control schools participated in the study. (See exhibit 1.6.) The resulting two study groups were similar in all but one of the 18 measures of school, principal, teacher, and student background characteristics examined: the percentage of principals with 4–10 years of experience. That percentage was lower for treatment principals than for control principals by a statistically significant amount (17 percent versus 33 percent). (See exhibits 1.7 to 1.10.)[36]

Exhibit 1.6. Random assignment results, fall 2012

Treatment status	Number of schools			Number of teachers	
	Total	Elementary schools	Middle schools	Elementary schools	Middle schools
Treatment	63	49	14	370	205
Control	64	48	16	366	228
Total	**127**	**97**	**30**	**736**	**433**

Exhibit Reads: There were 63 treatment schools in the study, of which 49 were elementary schools and 14 were middle schools. The treatment group elementary schools had 370 teachers, and the treatment group middle schools had 205 teachers.

[36] Separate baseline equivalence results for CLASS districts and FFT districts are provided in appendix A.

Exhibit 1.7. School background characteristics, by study group

Characteristic	Treatment group	Control group	Estimated difference	*p* value
Title I status (percentage)	69.8	73.2	-3.4	.448
Total school enrollment	511.0	513.7	-2.7	.865
Number of full-time equivalent teachers	32.1	31.9	0.2	.822
Percentage eligible for free and reduced-price lunch	40.0	40.8	-0.8	.565
Percentage minority	57.3	58.4	-1.0	.475
Percentage female	48.5	48.3	0.1	.759
Number of schools	**63**	**64**		

Exhibit Reads: Among the treatment group schools, 69.8 percent were Title I schools, and among the control schools, 73.2 percent were Title I schools. The estimated treatment and control group difference in the percentage of Title I schools was -3.4 percentage points, which was not statistically significant at the *p* < .05 level.

NOTE: The analyses are based on an OLS regression model controlling for random assignment blocks. The treatment group means are unadjusted means; the control group means were computed by subtracting the estimated group differences from the unadjusted treatment group means. The *p* values are based on *t* tests. Two-tailed statistical significance at the *p* < .05 level is indicated by an asterisk (*).

SOURCE: 2011–12 Common Core of Data.

Exhibit 1.8. Principal background characteristics, fall 2012, by study group

Characteristic	Treatment group	Control group	Estimated difference	*p* value
Years of experience in district				
Mean number of years	14.1	16.3	-2.2	.139
Three years or fewer (percentage)	19.0	8.6	10.4	.074
Four to 10 years (percentage)	17.5	33.2	-15.7*	.023
Eleven to 20 years (percentage)	33.3	25.7	7.7	.343
More than 20 years (percentage)	30.2	32.5	-2.3	.765
Master’s degree or higher (percentage)	†	†	-2.1	.480
Number of principals	**63**	**64**		

Exhibit Reads: Among the treatment group principals, the mean years of experience in the district was 14.1 years, and among the control group principals, the mean years of experience was 16.3 years. The estimated treatment and control group difference in mean years of experience was -2.2 years, which was not statistically significant at the *p* < .05 level.

NOTE: The analyses are based on an OLS regression model controlling for random assignment blocks. The treatment group means are unadjusted means; the control group means were computed by subtracting the estimated group differences from the unadjusted treatment group means. The *p* values are based on *t* tests. Two-tailed statistical significance at the *p* < .05 level is indicated by an asterisk (*).

† Figures suppressed due to small number of principals without a Master’s degree or higher.

SOURCE: Fall 2012 District Archival Records.

Exhibit 1.9. Teacher background characteristics, fall 2012, by study group (grades 4–8)

Characteristic	Treatment group	Control group	Estimated difference	*p* value
Years of experience in district				
Mean number of years	9.6	10.3	-0.7	.252
Three years or fewer (percentage)	25.8	24.8	1.0	.752
Four to 10 years (percentage)	37.9	34.8	3.0	.357
Eleven to 20 years (percentage)	23.9	25.4	-1.4	.597
More than 20 years (percentage)	12.3	14.8	-2.5	.308
Master's degree or higher (percentage)	43.9	46.1	-2.1	.396
Number of teachers	**575**	**594**		

Exhibit Reads: Among the grades 4–8 treatment group teachers, the mean years of experience in the district was 9.6 years, and among the control group teachers the mean years of experience was 10.3 years. The estimated treatment and control group difference in mean years of experience was -0.7 years, which was not statistically significant at the *p* < .05 level.

NOTE: The analyses are based on a two-level linear regression model controlling for random assignment blocks. The treatment group means are unadjusted means, and the control group means were computed by subtracting the estimated group differences from the unadjusted treatment group means. The *p* values are based on *t* tests. Two-tailed statistical significance at the *p* < .05 level is indicated by an asterisk (*).

SOURCE: Fall 2012 District Archival Records.

Exhibit 1.10. Student background characteristics, fall 2012, by study group (grades 4–8)

Characteristic	Treatment group	Control group	Estimated difference	*p* value
Students eligible for free or reduced-price lunch (percentage)	60.2	61.6	-1.4	.351
Race/ethnicity (percentage)				
White	44.2	43.1	1.1	.334
Black or African American	3.1	3.4	-0.3	.439
Hispanic	47.8	48.3	-0.6	.647
Asian/Pacific Islander	2.5	2.5	0.0	.991
Other	2.5	2.9	-0.4	.651
Female (percentage)	49.1	48.3	0.8	.204
English language learners (percentage)	15.6	16.9	-1.3	.360
Students with disabilities (percentage)	11.7	9.8	1.8	.159
Student achievement on state assessment (standardized)				
2011–12 Mathematics achievement	-0.009	-0.006	-0.003	.932
2011–12 Reading/ELA achievement	-0.029	0.022	-0.051	.111
Number of students	**15,551**	**17,308**		

Exhibit Reads: Among the treatment group students, 60.2 percent were eligible for free/reduced-price lunch, and among the control group students 61.6 percent were eligible. The estimated treatment and control group difference in percent eligible for free/reduced-price lunch was -1.4 percentage points, which was not statistically significant at the *p* < .05 level.

NOTE: The analyses are based on a three-level linear regression model controlling for random assignment blocks. The treatment group means are unadjusted means, and the control group means were computed by subtracting the estimated group differences from the unadjusted treatment group means. The *p* values are based on *t* tests. Two-tailed statistical significance at the *p* < .05 level is indicated by an asterisk (*).

SOURCE: District Archival Records.

Data Collection

Data for this report came from multiple sources as described next.

Data on districts' policies and practices. In spring 2013, the study team interviewed central office administrators about the districts' policies and practices related to performance feedback for all teachers and principals. The interviews also collected information about the integration of the study's intervention with existing district processes.

Data on the implementation of the intervention. We documented attendance at orientation and training events related to the study's performance measures. Online system records maintained by the organizations that supported the implementation of the performance measures and feedback were used for information on observer certification test pass rates, the frequency and timing of teacher observations and feedback sessions, and teachers' and principals' access of value-added reports. Surveys of observers hired by the study and interviews with district officials provided further information regarding the implementation of the observations and the district context, respectively.

Data on measures of educator performance. During the delivery of the intervention, data on measures of teacher classroom practice, student growth, and principal leadership were collected through the vendors' online systems. These data are for the treatment group only.

Data on educators' experiences with performance feedback. In spring 2013, we surveyed both the principals and teachers in all treatment and control schools. These surveys collected information on the nature and frequency of performance information educators received, and their perceptions of that information.

Data on the characteristics of study participants. To compare the characteristics of participants in the treatment and control groups, we collected data on school characteristics from the 2011–12 Common Core of Data and collected data on principals', teachers', and students' characteristics from district administrative records.

Each data collection resulted in a response rate of nearly 100 percent. The overall response rate for the teacher survey, for example, was 99.5 percent for grades K–8 teachers and 99.3 percent for grades 4–8 teachers; the response rate for the principal survey was 96.9 percent. More details about the types of data collected and data collection schedules are provided in appendix B.

Analytic Approaches

To examine the implementation of the intervention and the characteristics of the performance information educators received, we analyzed data collected from the treatment schools only. Specifically, to examine the implementation of the intervention, we conducted descriptive analyses of the extent to which study participants received the training on the performance measures, carried out the measurement activities, and received the performance information and feedback as planned.

To describe the characteristics of the performance information that teachers and principals received, we examined the distributions of scores (e.g., What percentage of principals had an overall rating of *distinguished*?) and the correlations among different performance measures. In addition, to estimate the reliability of the performance scores educators received, we used a generalizability theory framework (Shavelson and Webb 1991). Within this framework, reliability is defined as the proportion of variation in a measure's scores that reflect "true" differences between individuals rather than measurement error. The approach we used to define true versus error variation differed across the three measures based on the data available for each measure:

- We estimated the reliability of the teacher classroom practice ratings as a measure of the quality of stable classroom practice over a year, based on variation in ratings across the four observation windows.
- We estimated the reliability of the teacher value-added scores as a measure of stable teacher performance over the two years of student growth data that were used to calculate teacher value-added, based on the year-to-year variation in value-added scores.
- We estimated the reliability of the principal leadership ratings as a measure of leadership quality within each assessment window (fall and spring), based on variation in ratings across the three respondent groups.

(See appendix C for details about the reliability estimation methods.)

To assess whether the study's intervention led to differences in educators' experiences with performance feedback, we compared the survey responses of teachers and principals in the treatment and control groups, controlling for random assignment blocks. The specific analytic approach differed for binary survey measures (e.g., whether a teacher received feedback based on observations) and continuous survey measures (e.g., the number of instances of feedback received). For binary survey measures, we compared the mean for the treatment and control groups based on a principal-level linear probability model for principals and a two-level linear probability model (which accounted for the nesting of teachers within schools) for teachers. For continuous survey measures, we compared the median rather than the mean for the treatment and control groups because many of the survey-based continuous measures were not normally distributed.[37] (See appendix D for details about the analytic models.)

Organization of This Report

The remainder of this report includes four chapters. Chapters 2, 3, and 4 discuss the three performance measures that comprise the intervention (i.e., classroom practice, student achievement growth, and principal leadership), each providing an overview of the performance measure, findings about how well the performance measure and feedback were implemented, and findings about the performance ratings it generated. The last chapter (chapter 5) describes the contrasts between treatment and control teachers' and principals' experiences with performance feedback.

Additional information about the study and first-year findings is provided in appendixes A through I. Appendixes A and B present further details about the study sample and data collection. Appendixes C and D provide technical details about two sets of statistical analyses: reliability estimation and the assessment of the treatment-control differences in educators'

[37] The reported means and medians for the treatment group are unadjusted, and the means and medians for the control group were computed by subtracting the estimated group differences from the unadjusted treatment group means or medians.

experiences of performance feedback. Appendixes E, G, H, and I contain supplemental findings for the four findings chapters (i.e., chapters 2–5). Appendix F contains technical details about the estimation of value-added scores for the intervention's measure of student growth. The last appendix (appendix J) presents three sample reports: a CLASS observation report, an FFT observation report, and a student growth report for a principal.

Chapter 2. Findings About the Intervention’s Measure of Teacher Classroom Practice

This chapter presents findings about the intervention’s measure of teacher classroom practice as implemented during the first year of the study. As discussed in chapter 1, information provided by the measure is intended to distinguish teacher performance between lower- and higher-performing teachers, which could help identify educators in need of support, and to distinguish between different dimensions of a teacher’s practice, to highlight the dimensions on which an educator should improve.

The chapter begins by describing the design of the classroom practice measure and feedback. It then discusses findings about how well the measure and feedback were implemented, examines how well the measure differentiated teacher performance, and describes how well it identified practices for educators to improve. All findings in this chapter pertain to teachers in the treatment schools.

Key Findings

Implementation

- All observers were trained and certified.
- The majority (72 percent) of teachers received the intended four observations with feedback sessions.

Characteristics of the Classroom Practice Information

- For both CLASS and FFT, observation scores were concentrated at the upper end of the scale, limiting the degree of differentiation between lower- and higher-performing teachers.
- Teachers’ overall classroom observation scores, averaged across all four windows, contained measurement error, but provided some reliable information to distinguish between lower- and higher-performing teachers (reliability estimates between 0.42 and 0.75). The four-window average scores were positively but weakly correlated with teacher value-added scores (correlations of 0.09 for the CLASS and 0.17 for the FFT), and the correlation was statistically significant only for the FFT.
- Differences in a teacher’s ratings across observations limited how much one could learn about persistent performance from a single observation.
- Most (75 percent) of the CLASS observation reports identified at least one dimension of practice to improve and illustrated it with an example from the observation, but less than one quarter (23 percent) of FFT reports did so.

Overview of the Intervention’s Measure of Teacher Classroom Practice

The intervention’s measure of teacher classroom practice was designed to provide information on multiple dimensions of a teacher’s classroom practice repeatedly throughout the year. Specifically, it was designed to have the following features:

- four observations in each school year, one conducted by the principal or another school administrator and three conducted by study-hired observers[38];
- a report prepared by the observer after each observation, including ratings as well as narrative feedback; and
- an in-person feedback session after each observation during which the observer reviews the report with the teacher.

Districts were given the opportunity to choose between one of two distinct rating systems for measuring classroom practice. Four districts chose CLASS, and four districts chose FFT. The two systems capture similar dimensions of classroom practice but differ in how the observations and feedback sessions are conducted and in the amount and kind of information on teacher performance the systems' reports provide. The next section provides a description of each instrument.

The CLASS and FFT Rating Instruments and Observer Procedures

The CLASS-trained observers used the upper elementary version of CLASS, which is suitable for grades 4–8 teachers and covers 12 dimensions of classroom practice grouped into four domains (see exhibit 2.1).[39] Each CLASS observation includes two 25-minute cycles, which observers were asked to complete back-to-back. In each cycle, the observer spends the first 15 minutes observing the class and taking notes, and the next 10 minutes assigning a score to each dimension. All scores are on a 7-point scale. The observer then enters the dimension scores from each cycle into the CLASS online platform. The observer is also expected to write narrative text.

The FFT-trained observers used the Danielson Group's FFT, which is applicable across grades K–12. The observers focused on the two FFT domains that are observable, which together include 10 dimensions of classroom practice (see exhibit 2.1). Each FFT observation lasts at least 40 minutes. The observer takes notes during the observation period. Afterward, the observer organizes the notes in the FFT online system and enters scores for each dimension observed.[40] All scores are on a 4-point scale. For each scored dimension, the observer has the option to provide additional information for the report. Specifically, the observer may write narrative text relevant to that dimension and may select from a predefined list of teacher or student behaviors that indicate the basis for the dimension score. Finally, observers are required to write summary narrative text at the end of the report.

[38] In each treatment school, the classroom observations conducted by the principal or another school administrator were expected to be spread across the four observation windows. To the extent possible, each teacher was observed by the same study-hired observer over the school year, to build rapport with the teacher, which might improve the teacher's receptivity to the feedback. This was not always feasible, however, due to scheduling. Assigning these observations to different observers would have increased the reliability of the four-window average scores. However, we concluded that the potential benefits of rapport would outweigh the improved reliability.

[39] The different aspects of classroom practice are officially referred to as "dimensions" in the CLASS system and "components" in the FFT system. For simplicity, we use the term "dimensions" for both systems throughout this report.

[40] By design, not all FFT dimensions are necessarily observable during a classroom visit; therefore, FFT observers are not required to provide a score for each dimension for each observation.

Exhibit 2.1. Domains and dimensions of classroom practice for CLASS and FFT

Classroom Assessment Scoring System (CLASS-Upper Elementary)	Framework for Teaching (FFT)[a]
Domain 1: Emotional Support • Positive climate • Teacher sensitivity • Regard for student perspectives **Domain 2: Classroom Organization** • Behavior management • Productivity • Negative climate **Domain 3: Instructional Support** • Content development • Quality of feedback • Analysis and inquiry • Instructional dialogue • Instructional learning formats **Domain 4: Student Engagement** • Student engagement	**Domain 2: Classroom Environment** • Creating an environment of respect and rapport • Establishing a culture for learning • Managing classroom procedures • Managing student behavior • Organizing physical space **Domain 3: Instruction** • Communicating with students • Using questioning and discussion techniques • Engaging students in learning • Using assessment in instruction • Demonstrating flexibility and responsiveness

[a] The full FFT instrument includes two additional domains (Domain 1. Planning and Preparation, and Domain 4. Professional Responsibilities), which were not included as part of the intervention as they are not readily amenable to classroom observation.

Observer Procedures for Feedback Sessions for CLASS and FFT

Regardless of which instrument was used (i.e., CLASS or FFT), observers were required to notify teachers of the week when they would be observed. Thus, teachers knew approximately when they would be observed, although not the exact day and time. During the feedback sessions, the observers were to discuss two or three dimensions as the focus of the feedback session, including at least one strong dimension and one weak dimension. For each dimension, the observers were to talk about the behavioral indicators associated with the teacher's score and those associated with a higher score. Observers would then discuss actions the teacher could take to earn a higher score. To illustrate the teaching practices that were discussed, CLASS observers were to show the teacher one or two videos relevant to the focal dimensions and recommend additional videos for the teacher to view on his or her own. FFT observers were to recommend videos and other resources on the Teachscape website that the teacher could review for help thinking about how to improve his or her instruction.

The CLASS and FFT Reports

The CLASS and FFT online platforms were designed to provide each teacher with a report on each observation, which the observer would review with the teacher during the in-person feedback session after the observation. The observation reports generated by the online platforms differ in content, with the CLASS reports providing more scores, performance levels tied to scores, and more narrative text.

The CLASS reports are organized hierarchically, providing three levels of scores:

- one overall score, which is the average of the 12 dimension scores;
- four domain scores, each of which is an average of the dimension scores within the domain;[41] and
- 12 dimension scores, each of which is an average of the dimension scores from the two cycles.

Each score is accompanied by one of four performance levels: *ineffective*, *developing effectiveness*, *effective*, and *highly effective*. For the overall score and each domain score, the report additionally presents a graphic comparison of the teacher's score with the district average for each of the previous observations.[42]

In addition to the scores, the CLASS reports provide narrative text for the overall lesson**,** as well as for each domain and dimension. The report text begins with a narrative description of the lesson (i.e., "Context of the Observation") and a summary of how the teacher performed overall (i.e., "CLASS Advisor Summary"). Then, for each domain, the reports provide a narrative summary of how well the teacher performed on that domain. Finally, for each dimension, the reports describe examples of "effective" and "less effective" behaviors observed. See appendix J for a sample CLASS report.

Unlike the CLASS reports, the FFT reports present only the score for each observed dimension; the FFT reports do not present an overall score or domain scores. In addition, although teachers may remember the performance levels associated with each score from the orientation, described in chapter 1, the FFT reports do not mention performance levels unless the observer uses an optional feature of the online system to do so.[43] The performance levels are *unsatisfactory, basic, proficient,* and *distinguished*. See appendix J for a sample FFT report.

The CLASS and FFT online platforms were also equipped to provide each principal with reports on all of the teachers he or she supervises. To ensure that building administrators received the performance information for all teachers in their respective schools, the study held midyear meetings of the building administrators in each district in the winter to review classroom practice reports and learn how to access reports on the online platform.

[41] The Student Engagement Domain includes only one dimension, as shown in exhibit 2.1; therefore, for this domain, the domain score is the same as the dimension score.

[42] The district average refers to the average across all of the treatment schools in the district.

[43] When completing an FFT report, an observer could select from a predefined list of teacher or student behaviors that each correspond to a performance level. When a behavior is selected, it appears in the report under the relevant dimension in a special subsection called "critical attributes," along with the performance level. For example, an observer could select "Proficient—Teacher responds to disrespectful behavior among students," and those exact words would appear on the report. In an analysis of a random sample of FFT reports from year 1, we found that 48 percent were not missing "critical attributes" or were missing them for only one or two of the scored dimensions.

Findings About the Implementation of the Intervention's Measure of Classroom Practice

To assist with implementation of the performance measures tested in this study, an AIR team separate from the evaluation team monitored implementation and provided support when needed to keep the activities on track (e.g., to ensure that most teachers were observed approximately four times per year). The implementation team also worked with each district to identify and hire observers to conduct the observations and feedback sessions consistent with the study design. To ensure that observers had experience giving feedback, districts were asked to identify teachers with prior experience with coaching or mentoring. Observers received the standard training offered by the CLASS and FFT vendors, which is designed to teach observers to reliably score instruction. In addition, observers received training in how to enter scores and narrative for the classroom observation reports and how to conduct the feedback sessions with teachers. These training components were developed by the CLASS and FFT vendors for the study based on existing materials.

This section presents findings about the extent to which the intervention's measures of teacher classroom practice were implemented as intended, focusing in particular on observer training and certification, observers' perceptions of the systems, the number of observations and feedback sessions that teachers received, and teacher engagement during the feedback sessions. Results are presented for the full sample as well as for the CLASS and FFT districts separately. Because districts were not randomly assigned to the CLASS or FFT version of the intervention, differences in the results in the CLASS and FFT districts cannot necessarily be attributed to the CLASS and FFT observation systems. Any differences might be due to other district characteristics.

Observer Training and Certification

All observers were trained and passed the certification test, often after multiple attempts. According to the vendors' online system records, all school administrators and study-hired observers who conducted observations attended the training on how to score classroom practice and provide written and oral feedback to teachers; nearly all (92 percent for CLASS and 97 percent for FFT) completed all of the required training, which lasted three days for CLASS and four days for FFT.[44] All observers passed the certification test, which included video scoring exercises. However, it took multiple attempts to pass the test for half of the CLASS observers and 17 percent of the FFT observers.[45]

Of the study-hired observers who were certified, nearly all (94 percent for CLASS and 99 percent for FFT) reported that they had at least five years of teaching experience, as intended, and nearly all (95 percent for CLASS and 92 percent for FFT) also had experience supervising others or working as an instructional coach. (See exhibit A.8 in appendix A for further details

[44] Among those who did not complete all of the required training, the mean number of days of training received was 1.6 for CLASS and 2.6 for FFT.

[45] The CLASS certification test allowed up to four attempts to pass. The FFT certification test included two stages, and observers could attempt each stage three times. The implementation team monitored observers' progress in becoming certified and offered additional training to observers who didn't pass.

about the background characteristics of study-hired observers.) The certified study-hired observers were assigned to teachers primarily based on geography rather than subject area or grade.

The majority of observers reported positive perceptions of the rating systems. Survey items were included on the principal survey and the survey of study-hired observers to determine whether the observers bought into the rating systems. Based on survey data, nearly all principals agreed somewhat or strongly that the CLASS/FFT rating system accurately reflected the quality of an individual's teaching (92 percent), did a good job distinguishing effective from ineffective teaching (97 percent), and was fair to all teachers (95 percent). Most study-hired observers also expressed positive views of the rating systems (93 percent, 88 percent, and 80 percent, respectively). (See exhibit E.1 in appendix E for the results for CLASS and FFT separately.)

Classroom Observations and Feedback Sessions

The majority of teachers received the intended four observations with feedback sessions. According to the online system records maintained by the vendors, a large majority (82 percent) of teachers received all four observations and nearly three quarters (72 percent) received all four feedback sessions (see exhibit 2.2).[46] On average, each teacher received 3.8 observations (3.7 for CLASS and 3.9 for FFT) and 3.7 feedback sessions (3.5 for CLASS and 3.9 for FFT) during the first year of the study.

Approximately two thirds of teachers were observed once by a school administrator and three times by one or more study-hired observers, as intended. The intervention's measures of classroom practice were designed to provide teachers with observations conducted by their school administrators, who should know the teachers well, and observations conducted by observers from outside the schools, who might be more impartial in their assessment of teacher performance.[47] Approximately two thirds (66 percent) of teachers (49 percent for CLASS and 90 percent for FFT) were observed once by a school administrator (typically the principal) and three times by one or more study-hired observers, as intended. On average, each teacher was observed by 2.5 different observers (2.7 for CLASS and 2.3 for FFT) during the year.

[46] Some teachers did not receive all four observations or feedback sessions for reasons such as maternity leave or medical leave. Seven teachers received five observations during the year because they were observed twice by two different study-hired observers during the first observation window due to scheduling confusion. For the analysis presented in exhibit 2.2, we classified these seven teachers as receiving four observations during the year.

[47] When assigning observers to teachers, districts were asked to minimize assigning observers to teachers they already knew. To examine whether the study-hired observers had a pre-existing relationship with any of the teachers they observed, we asked them, "Of your assigned teachers, how many did you already know on a personal or professional basis before your TLES observation work began." On average, study-hired observers reported that they had already known 5 percent on a personal basis and 19 percent on a professional basis. Also, none of the study-hired observers worked in a school-based role during the intervention years. Rather, the study-hired observers were all former teachers, coaches, or administrators living in or around the school districts, or current employees of the central office (for more detail, see appendix exhibit A.8).

Exhibit 2.2. Percentage of teachers who received one, two, three, or four study observations and feedback sessions in CLASS and FFT districts

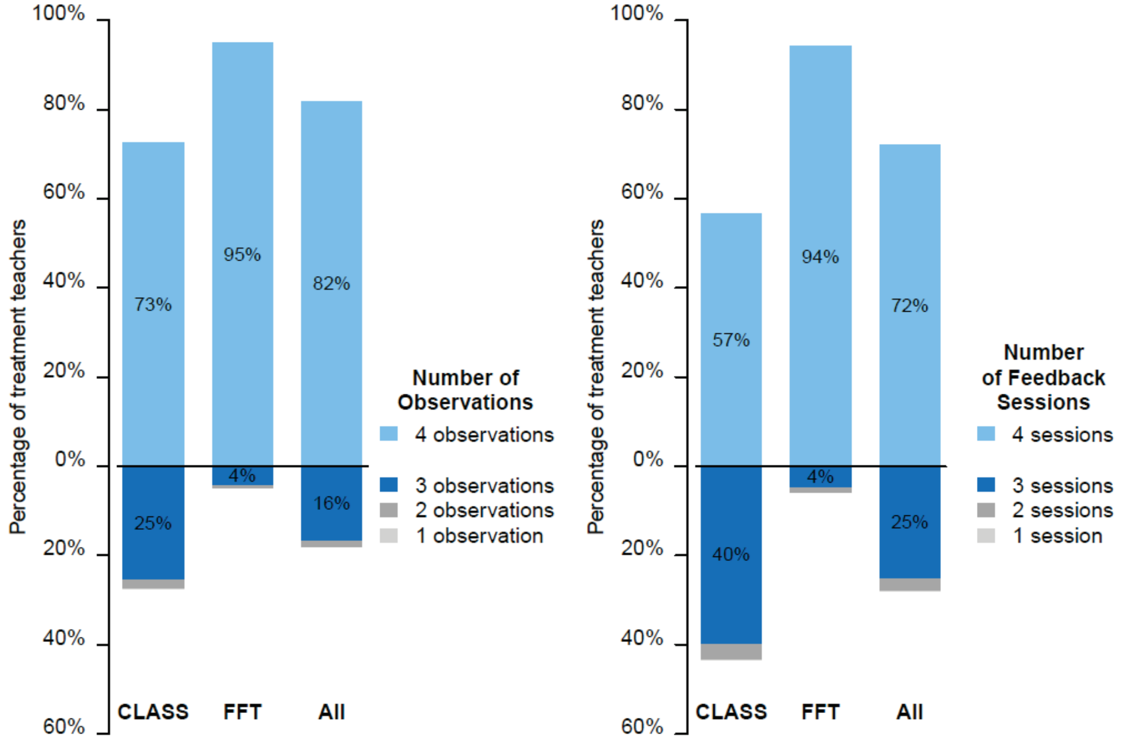


Exhibit Reads: Of the treatment teachers in CLASS districts, 72.5 percent received four observations during the year, 25.2 percent received three observations, less than 3 percent received two observations, and no teacher received one observation. Of the CLASS treatment teachers, 56.5 percent received four feedback sessions, 39.6 percent received three feedback sessions, less than 4 percent received two feedback sessions, and no teachers received one feedback session.

NOTE: Sample size = 535 teachers (313 CLASS and 222 FFT). See exhibit E.2 in appendix E for results for K–3 teachers.

SOURCE: Teachstone Online System and Teachscape Online System.

The majority of study-hired observers reported that teachers appeared interested and engaged during the feedback sessions. Nearly all (96 percent) of the study-hired observers in CLASS districts reported that teachers seemed genuinely interested in using the feedback to improve their teaching in more than two thirds of the feedback sessions. Only two thirds of the study-hired observers in FFT districts, however, reported the same. Across all districts, the majority of study-hired observers (79 percent for CLASS and 81 percent for FFT) reported that teachers were actively engaged in discussions in more than two thirds of the feedback sessions.

Findings About the Classroom Practice Information

As described previously in the chapter, the classroom observation measure included detailed information for teachers on their teaching. Such information was conveyed both in writing, through a report created by the observer using the online system, and in a feedback session with the observer after each observation. The CLASS reports included scores and corresponding performance levels at the dimension level, domain level, and overall. The FFT reports included scores at the dimension level only. For analytic purposes, the study's evaluation team created an overall score for each FFT observation by averaging the ten FFT dimension scores, each of which was on a 1 to 4 scale. These overall scores were rounded to the nearest whole number to create four study-defined performance levels aligned with the FFT dimension scores and the corresponding performance levels (e.g., 1 corresponds to *unsatisfactory*).

The study data do not allow us to determine specifically what information, if any, teachers focused on. For analyses in the following subsection, we look at overall scores within each observation window and average across the four observation windows.[48] For analyses in the subsequent subsection, we look at scores at the dimension level for each observation window. Finally, although we present these results for CLASS and FFT districts separately, it is important to keep in mind that differences in the results in the CLASS and FFT districts cannot necessarily be attributed to the CLASS and FFT observation systems. Because districts were not randomly assigned to the CLASS or FFT version of the intervention, differences could occur due to other district characteristics.

Variation in Teacher Performance Based on Overall Ratings

A key question is whether the classroom practice ratings as implemented in the study districts distinguished between teachers whose persistent performance during the year was better or worse. To determine whether the ratings distinguished between teachers, we looked at how the scores varied from window to window and whether they provided a reliable measure of a teacher's persistent classroom practice during the year. Finally, to check the validity of the classroom practice measures as implemented in the study, we looked at the relationship between the observation scores and student learning as measured by value-added scores.

Nearly all teachers had classroom observation overall scores in the top two performance levels, limiting the degree of differentiation between lower- and higher-performing teachers. Within each observation window, for CLASS observations, nearly all (98 percent) of the teachers received an overall score that placed them in the top two performance levels, labeling them *effective* or *highly effective* (see exhibit 2.3). The percentage of CLASS teachers labeled as *highly effective* increased over the four observation windows from 74 percent in the first window to 89 percent in the fourth window. Less than 2 percent of the CLASS teachers within an observation window received an overall score that placed them in the bottom two performance levels, and no teachers had a four-window average overall score in the bottom two performance levels.

For FFT, more than 88 percent of the teachers within an observation window had an overall score of 2.50 or higher, which corresponds to the top two study-defined performance levels (see exhibit 2.4). More than three quarters of the teachers had overall scores between 2.50 and 3.49, which corresponds to the second highest performance level. Within each observation window, less than 12 percent of the FFT teachers received an overall score less than 2.50, which

[48] The "four-window average" overall score represents the average overall score a teacher received during the year. For most teachers, this average score is based on overall scores from each of the four observation windows. Some teachers had fewer than four observations (see exhibit 2.2), so the average score is based on the number of observations they had during the year.

corresponds to the bottom two study-defined performance levels, and 5 percent of teachers had a four-window average overall score that was less than 2.50.

Exhibit 2.3. Distribution of teachers across performance levels based on CLASS overall scores, by observation window and the four-window average

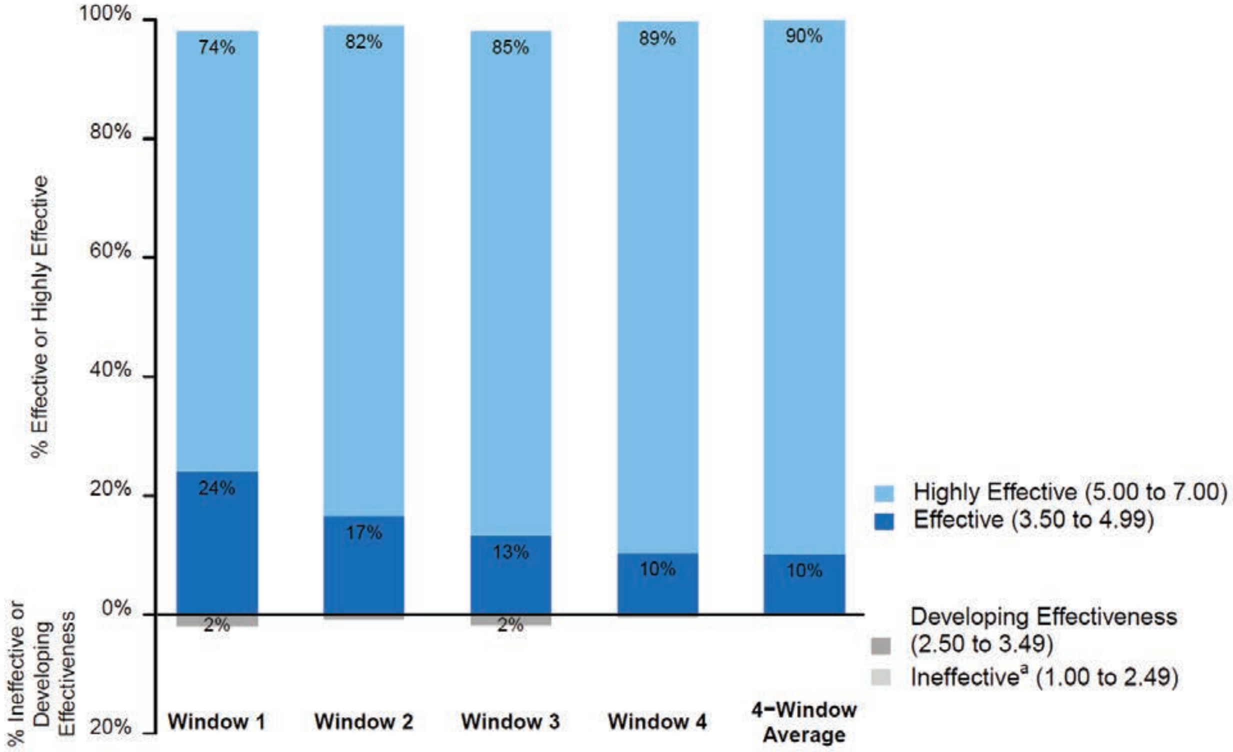


Exhibit Reads: Of treatment teachers in CLASS districts observed in window 1, 74 percent had a CLASS overall score at the *highly effective* performance level, 24 percent at the *effective* performance level, and 2 percent at the *developing effectiveness* performance level. Less than 1 percent of teachers had an overall score at the *ineffective* performance level.

NOTE: Sample size = 262 teachers in window 1, 307 teachers in window 2, 309 teachers in window 3, 279 teachers in window 4, and 313 teachers for the four-window average. Reported percentages may not sum to 100 percent because of rounding.

[a] Within a window, less than 1 percent of teachers had an overall score at the *ineffective* performance level.

SOURCE: Teachstone Online System.

Exhibit 2.4. Distribution of teachers across study-defined performance levels based on FFT overall scores, by observation window and the four-window average

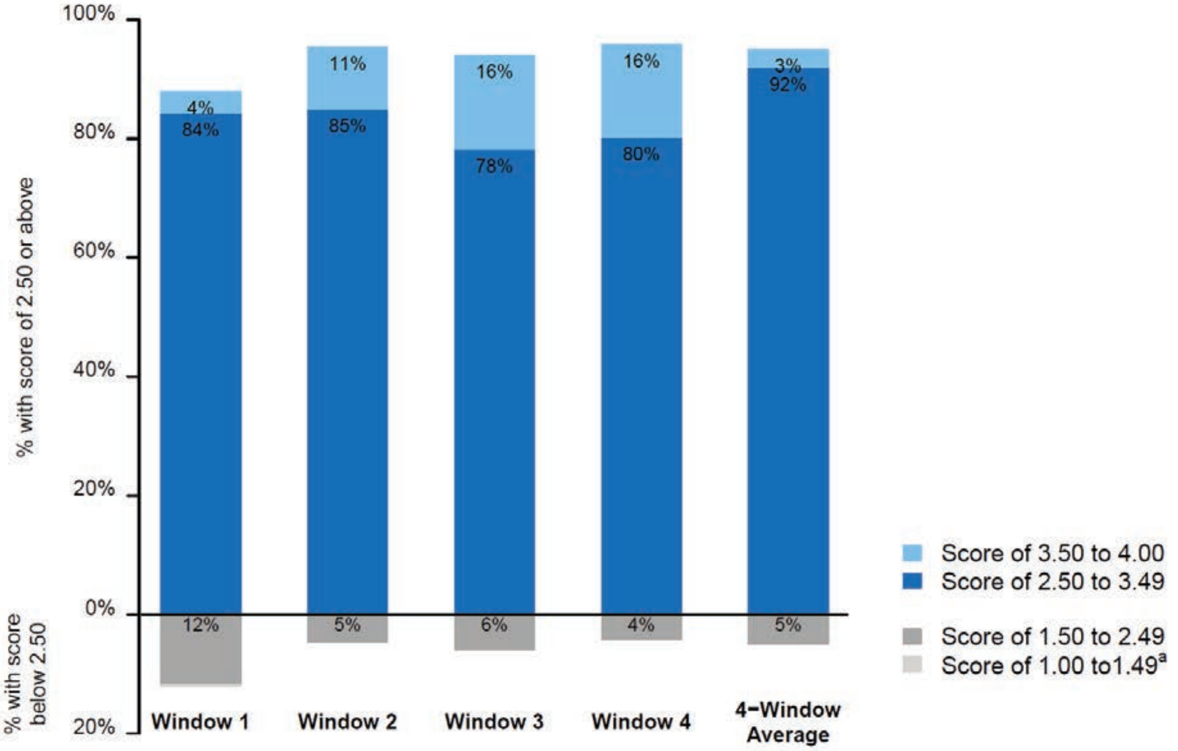


Exhibit Reads: Of treatment teachers in FFT districts observed in window 1, 4 percent had an FFT overall score between 3.50 and 4.00, 84 percent had a score between 2.50 and 3.49, and 12 percent had a score between 1.50 and 2.49. Less than one percent of teachers had an overall score below 1.50.

NOTE: The distribution in each window is based on teachers' FFT overall scores categorized into study-defined performance levels. To create the overall scores and performance levels, the study's evaluation team first calculated an overall score by averaging the teacher's ten FFT dimension scores, each of which was on a 1 to 4 scale. The overall scores were then categorized into study-defined performance levels by rounding them to the nearest whole number. This created four performance levels aligned with the FFT dimension scores. An FFT dimension score of 1 corresponds to *unsatisfactory*, 2 corresponds to *basic*, 3 corresponds to *proficient*, and 4 corresponds to *distinguished*. Average FFT scores and overall performance levels are not provided in the FFT reports teachers received. Sample size = 216 teachers in window 1, 219 teachers in window 2, 220 teachers in window 3, and 217 teachers in window 4. Reported percentages may not sum to 100 percent because of rounding.

[a] Within a window, less than 1 percent of teachers had an overall score below 1.50.

SOURCE: Teachscape Online System.

Classroom observation overall scores were concentrated toward the high end of the rating scale but still varied across teachers. The overall score distributions indicate that there were differences in teachers' overall scores, even among teachers with the same performance level designation. The distributions of CLASS and FFT overall scores in each window and the averaged scores across the four windows are presented in exhibits 2.5 and 2.6, respectively. Each exhibit shows that although the scores are spread primarily across the upper half of the rating scales, there is still variation.[49]

[49] For CLASS, the average overall score teachers were assigned by study-hired observers was 0.28 points higher than the average score they were assigned by their principal ($p < .05$). For FFT, the average overall score teachers

Exhibit 2.5. Distribution of teachers based on their CLASS overall scores in each observation window and the four-window average

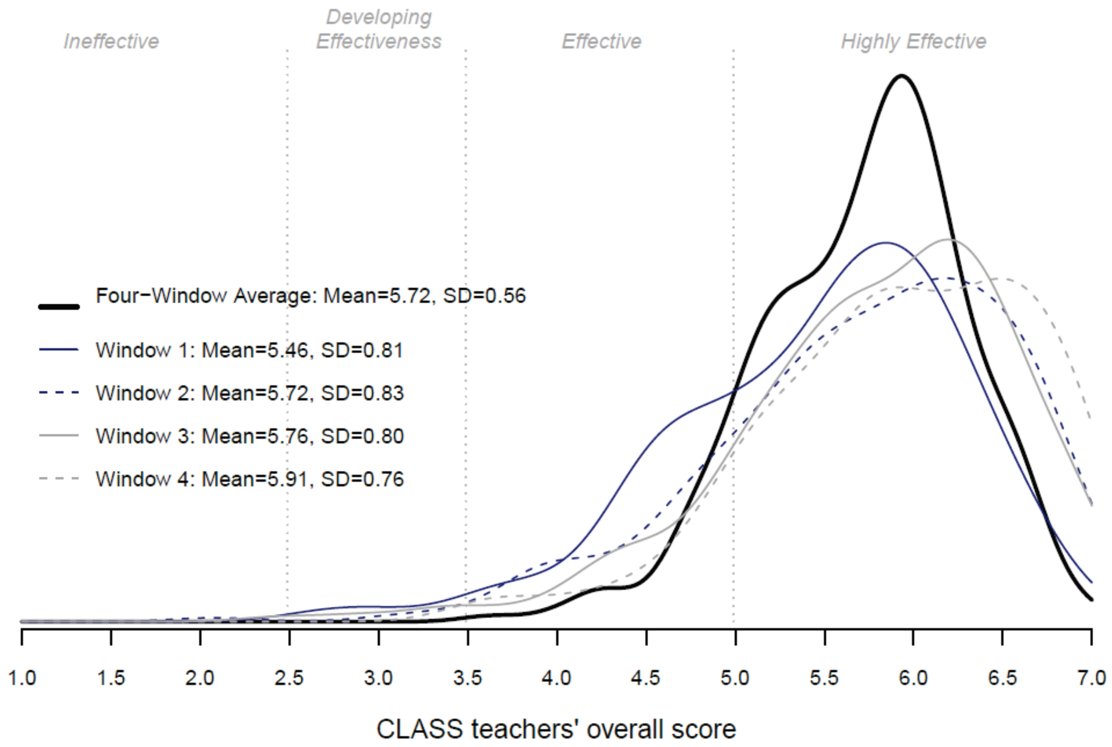


Exhibit Reads: The four-window average CLASS overall scores were concentrated within the *highly effective* performance level, with a score just below 6.0 being the most common four-window average overall score.

NOTE: The exhibit shows the density of teachers across the score distribution, where the area under each curve between two scores represents the percentage of teachers with scores in that range, and the total area under the curve sums to 100 percent. Sample size = 262 teachers in window 1, 307 teachers in window 2, 309 teachers in window 3, 279 teachers in window 4, and 313 teachers for the four-window average. See exhibit E.9 in appendix E for detailed information about the distribution of four-window average CLASS observation scores for K–3 teachers.

SOURCE: Teachstone Online System.

were assigned by study-hired observers was not statistically significantly different from the average overall scores teachers were assigned by their principal (p = .111).

Descriptive statistics for the four-window average scores are presented in exhibits E.3 and E.4 in appendix E. Mean four-window average overall scores by teacher characteristics are presented in exhibits E.5 and E.6 in appendix E, and average scores by observer type (study-hired observers and principals) are presented in exhibits E.7 and E.8.

Exhibit 2.6. Distribution of teachers based on their FFT overall scores in each observation window and the four-window average

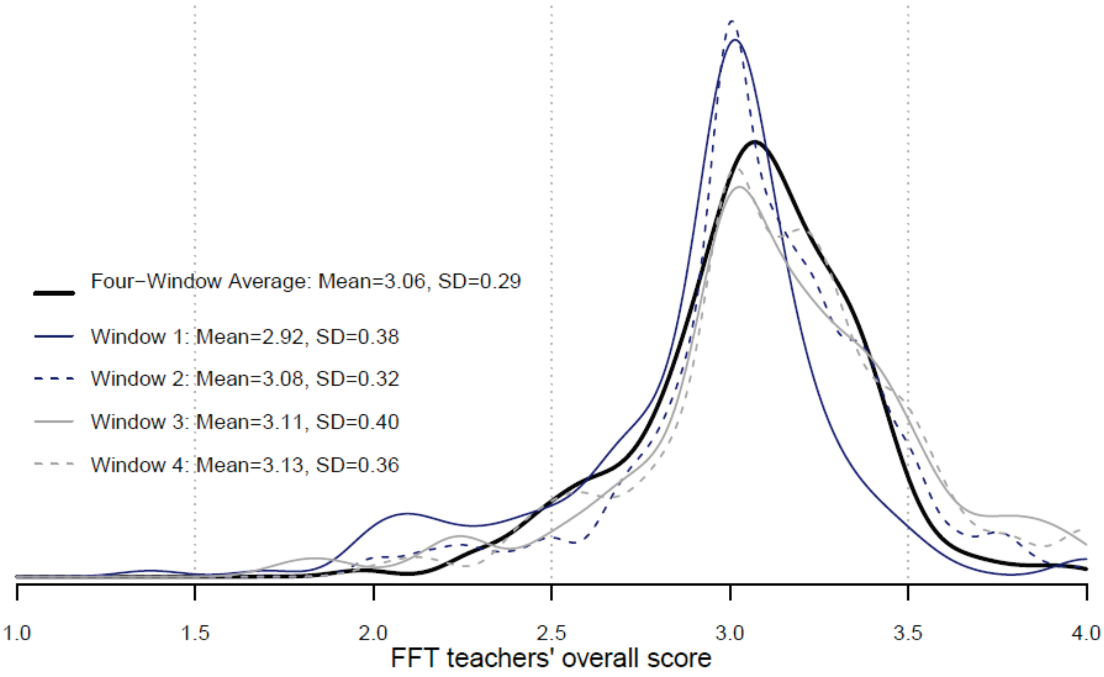


Exhibit Reads: The four-window average FFT overall scores were concentrated between a score of about 2.5 and 3.5, with a score just above 3.0 being the most common four-window average overall score.

NOTE: The exhibit shows the density of teachers across the score distribution, where the area under each curve between two scores represents the percentage of teachers with scores in that range, and the total area under the curve sums to 100 percent. The grey dotted vertical lines represent cut-points for the study-defined performance levels. Average FFT scores and overall performance levels were not provided in the FFT reports teachers received. Sample size = 216 teachers in window 1, 219 teachers in window 2, 220 teachers in window 3, 217 teachers in window 4, and 222 teachers for the four-window average. See exhibit E.10 in appendix E for detailed information about the distribution of four-window average FFT observation scores for K–3 teachers.

SOURCE: Teachscape Online System.

The overall score averaged across four windows provided some reliable information to distinguish between lower- and higher-performing teachers, but differences in a teacher's ratings across observations limited how much one could learn about persistent performance from a single observation. To distinguish between lower- and higher-performing teachers, the CLASS and FFT overall scores need to provide a reliable way for teachers and principals to identify patterns in performance over the course of the year. This means a teacher's overall scores should reflect persistent classroom practice over the year and not mainly idiosyncratic factors introduced by the observer or the particular days or lessons observed.[50] We estimated the degree to which the overall scores were a reliable measure of a teacher's persistent classroom practice over the year based on how much a teacher's overall scores varied across the four observation windows relative to how much the

[50] Classroom practice ratings from a single observation could also inform feedback about a teacher's instruction during a particular lesson, even if that performance is 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.

four-window average scores varied across teachers (see appendix C for details about the estimation methods and results). These reliability estimates tell us how consistent a teacher's overall scores were over the four observation windows. Educators should have more confidence in decisions and actions based on more reliable measures, though what constitutes "sufficient" reliability depends on the measure's intended use.[51] The analyses produced the following findings[52]:

- The four-window average overall scores contained measurement error but provided some reliable information about a teacher's classroom practice over the year. Depending on assumptions about the sources of variation, reliability estimates for the four-window average overall scores were between .42 and .50 for CLASS and between .69 and .75 for FFT.
- Overall scores based on a single observation had limited reliability as a measure of a teacher's persistent classroom practice over the year because of variation in a teacher's overall scores across the four observation windows. The reliability of overall scores based on a single observation was .24 for CLASS and .49 for FFT, which is consistent with the correlations of the overall scores across windows (see exhibit E.11 in appendix E). In other words, 24 percent of the variation in CLASS overall scores and 49 percent of the variation in FFT overall scores represented between-teacher differences in classroom practice.

Classroom observation overall scores were positively, although weakly, associated with teacher value-added scores. If the classroom observation scores were a valid measure of teacher classroom practice, then a teacher's observation score should be associated with other indicators of the quality of classroom practice. To check this, we examined the correlation of teachers' overall observation scores with their value-added scores from the prior year.[53] For both CLASS and FFT, the four-window average overall scores had a positive relationship with teachers' prior-year value-added scores (see exhibit 2.7). The correlation between the four-window average overall scores and the prior-year value-added overall scores was .09 for CLASS and .17 for FFT. For CLASS, the observation scores had a stronger relationship with reading/ELA value-added scores than with mathematics value-added scores. For FFT, the observation scores had a similar relationship with value-added scores in

[51] The Standards for Educational and Psychological Testing (AERA/APA/NCME 2014) do not suggest a minimal degree of reliability, but state that the reliability evidence for a measure should be appropriate for the measure's intended use, and a higher degree of reliability is required for uses that have more significant consequences. For consequential personnel decisions, measures with reliabilities above .70 are often considered acceptable (U.S. Department of Labor 2006), though job performance ratings have been found to often have reliabilities below .70 (Viswesvaran, Ones, and Schmidt 1996).

[52] The reliability estimates are consistent with findings from other studies of classroom observation reliability (Casabianca et al. 2013; Ho and Kane 2013; Kane and Staiger 2012). For example, Casabianca et al. (2013) report reliabilities for CLASS that range from .32 to .72, and Ho and Kane (2013) report a reliability for FFT of .66.

[53] The correlations between observation scores and prior-year value-added scores is more informative than the correlation between observation scores and current-year value-added scores because the relationships between observation scores and value-added scores based on the same classroom of students can have correlated error terms, which may artificially inflate measures of association (Kane and Staiger 2012). These cross-year correlations may, however, underestimate the true correlation if teachers' performance changed appreciably from one year to the next. For comparison, same-year correlations are provided in Exhibit E.15. As expected, correlations with current-year value added are stronger than correlations with prior-year value-added.

reading/ELA and mathematics. Although the correlations between classroom observation overall scores and value-added scores were modest in magnitude, these correlations are consistent with the magnitudes found by other studies (Chaplin et al. 2014; Kane and Staiger 2012; Kane et al. 2011) and likely underestimate the strength of the true association because of measurement error in both the observation scores and the value-added scores. Correlations based on the window-specific overall scores were generally lower than correlations based on the four-window average overall scores, which reflects the limited reliability of the window-specific overall scores.

Exhibit 2.7. Pairwise correlations between classroom observation overall scores and prior-year value-added scores

	Overall[a]		Mathematics		ELA/Reading	
	N	Correlation coefficient	*N*	Correlation coefficient	*N*	Correlation coefficient
CLASS						
Four-window average	253	.09	198	.04	182	.17*
Window 1	217	.07	170	.05	156	.14
Window 2	251	.11	196	.09	180	.10
Window 3	252	.11	197	.08	182	.20*
Window 4	226	.00	186	-.04	166	.09
FFT						
Four-window average	173	.17*	142	.21*	142	.15
Window 1	169	.15	138	.16	139	.14
Window 2	171	.09	140	.12	140	.13
Window 3	173	.17*	142	.19*	142	.17*
Window 4	171	.10	141	.15	140	.07

Exhibit Reads: The correlation between the four-window average CLASS overall scores and teachers' prior year overall value-added scores was 0.09 based on 253 treatment teachers in CLASS districts.

[a] The overall value-added score for a teacher with value-added scores in both mathematics and reading/ELA is a precision-weighted average of the value-added scores in both subjects. The overall value-added score is the same as the subject-specific value-added score for teachers with a value-added score in only one subject.

Two-tailed statistical significance at the $p < .05$ level is indicated by an asterisk (*).

SOURCE: Teachstone Online System (CLASS), Teachscape Online System (FFT), and Student Growth Reporting System.

Feedback on Specific Dimensions of Classroom Practice

To inform decisions regarding the focus of professional development, the classroom observation reports provide performance information on different dimensions of classroom practice. The reports additionally allow observers to justify scores by describing teachers' specific practices in narrative form. We begin this section by looking at the variation in scores across the various dimensions. The remainder of the section presents findings on the extent to which the classroom observation reports provided teachers with reliable information at the dimension level to guide decisions about the focus of professional development.

Most teachers received classroom observation scores that differed across dimensions of classroom practice. Both CLASS and FFT reports provided teachers with separate scores for different dimensions of their classroom practice. If a teacher received different scores on different dimensions of their classroom practice, then that might allow the teacher to see on which dimensions of classroom practice he or she performed relatively well and on which he or she performed relatively poorly.

An analysis of the extent to which teachers' scores spanned one, two, three, or four performance levels produced the following findings:

- In CLASS districts, teachers received observation scores for 12 dimensions of classroom practice (see exhibit E.12 in appendix E for their correlations). In the first two observation windows, more than half of the CLASS teachers (61 percent and 52 percent, respectively) received scores at multiple performance levels, and just under half did so in the last two observation windows (49 percent and 42 percent, respectively) (see exhibit 2.8).
- In FFT districts, teachers received observation scores for up to 10 dimensions of classroom practice (see exhibit E.13 in appendix E for their correlations). In each observation window, more than two thirds of the FFT teachers received scores in multiple performance levels (see exhibit 2.9).

Exhibit 2.8. Percentage of teachers whose CLASS dimension scores spanned one, two, three, or four performance levels, by observation window

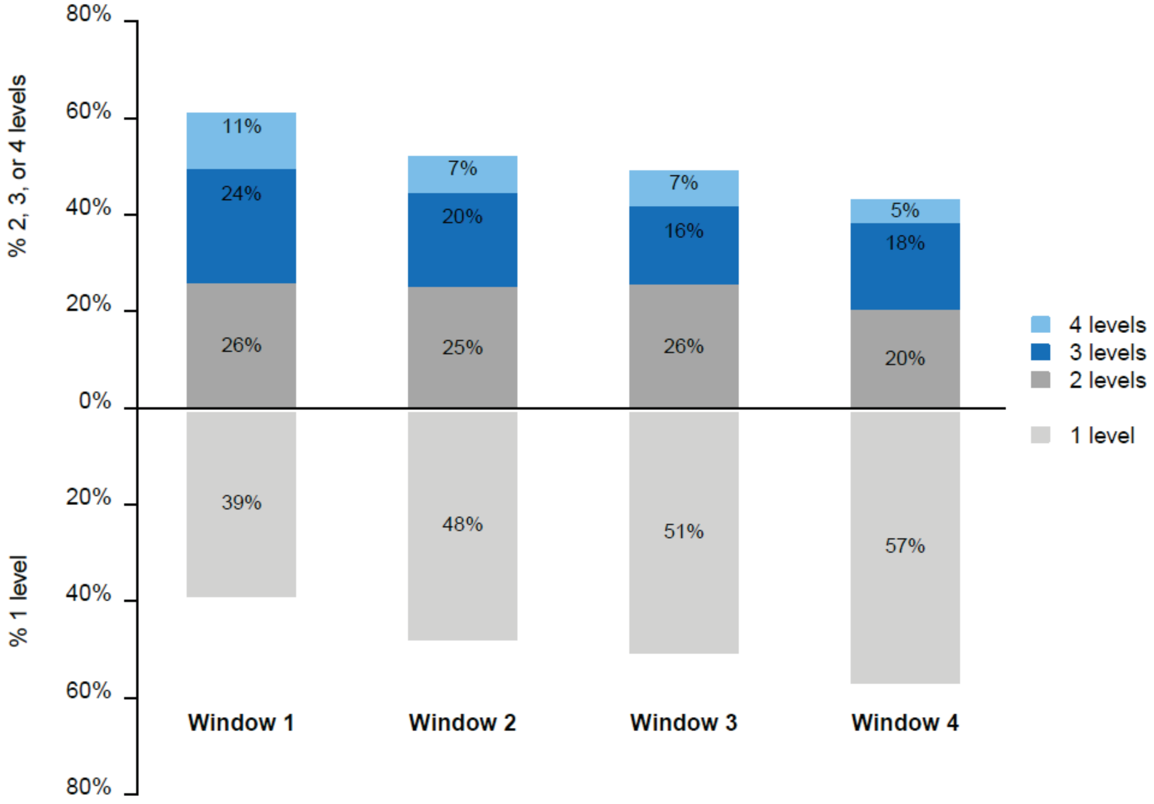


Exhibit Reads: Of treatment teachers in CLASS districts observed in window 1, 11 percent had dimension scores that fell into four different performance levels, 24 percent had dimension scores in three different performance levels, 26 percent had dimension scores in two different performance levels, and 39 percent had all dimension scores in the same performance level.

NOTE: Sample size = 262 teachers in window 1, 307 teachers in window 2, 309 teachers in window 3, and 279 teachers in window 4. Reported percentages may not sum to 100 percent because of rounding.

SOURCE: Teachstone Online System.

Exhibit 2.9. Percentage of teachers whose FFT dimension scores spanned one, two, three, or four performance levels, by observation window

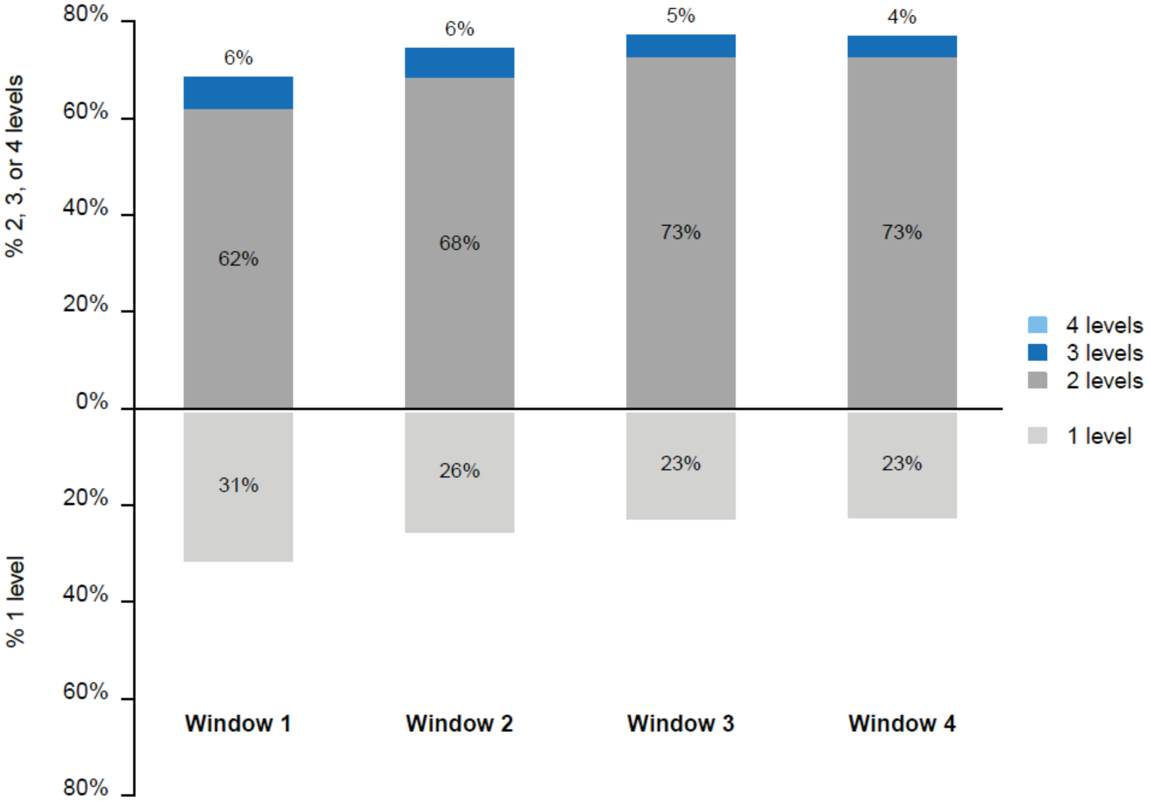


Exhibit Reads: Of treatment teachers in FFT districts observed in window 1, 6 percent had dimension scores in three different performance levels, 62 percent had dimension scores in two different performance levels, and 31 percent had all dimension scores in the same performance level (no teacher had dimension scores in four different performance levels).

NOTE: The different aspects of classroom practice are officially referred to as “dimensions” in the CLASS system and “components” in the FFT system. For simplicity, we use the term “dimensions” for both systems. Sample size = 216 teachers in window 1, 219 teachers in window 2, 220 teachers in window 3, and 217 teachers in window 4. Reported percentages may not sum to 100 percent because of rounding.

SOURCE: Teachscape Online System.

Although many teachers received scores that spanned multiple levels, the scores did not reliably distinguish between different dimensions of a teacher’s classroom practice. Even though many teachers received scores that spanned multiple performance levels, a teacher’s scores may not have been useful if the scores did not convey a consistent message over the year about the teacher’s relative performance across dimensions of classroom practice. For example, less than a third of the teachers (12 percent for CLASS and 27 percent for FFT) had the same lowest-scored dimension of classroom practice in each of the four observation windows. We estimated the degree to which a teacher’s scores over the year were a reliable measure of persistent differences in a teacher’s relative performance between dimensions of classroom practice based on how much teachers’ scores differed relative to differences in scores over the four observation windows (see appendix C for details about the estimation methods and results).

The analysis produced the following findings:

- The estimated reliability of the difference between the observation scores for two different dimensions of classroom practice was only .19 for CLASS and .09 for FFT based on scores from a single observation window.
- Differences based on the four-window average dimension scores were more reliable than dimension scores based on a single window but were still limited by measurement error. Depending on assumptions about the sources of variance, reliability estimates for the difference between a teacher's four-window average dimension scores for different dimensions of classroom practice were between .35 and .43 for CLASS and between .18 and .23 for FFT.

Most of the CLASS observation reports identified at least one dimension of classroom practice to improve and illustrated it with an example from the observation, but fewer than a quarter of FFT reports did so. The observers wrote narrative text identifying at least one dimension of practice as a strength and one dimension for improvement, as required, in the majority of the observation reports (76 percent of CLASS reports and 71 percent of FFT reports, based on an analysis of 160 randomly selected reports). Three quarters of the sampled CLASS reports supported the identified dimension(s) of practice for improvement with at least one example from the observation, but less than a quarter (23 percent) of the sampled FFT reports did so. This difference might be at least partly attributable to the difference in reporting requirements between CLASS (which required the observers to fill out all fields) and FFT (which did not require the observers to fill out all dimension-specific fields).

Summary

Study districts were largely successful in implementing the teacher classroom practice performance measure. Observers were trained and certified, and they completed four rounds of observation and feedback for most teachers, as intended. However, the performance information did not fully distinguish teacher performance. On the one hand, teachers' overall observation scores averaged across the four windows were positively correlated with their value-added scores, and the scores differentiated performance, identifying lower- and higher-performing teachers. Between-teacher variation in these scores was more than would be expected by chance. On the other hand, these four-window average scores only appeared in the CLASS reports provided to teachers; they did not appear in the reports provided by the FFT. In addition, for both CLASS and FFT, the scores were concentrated at the upper end of the scale and virtually all teachers had scores associated with positive performance levels (e.g., *effective* and *highly effective*). More detailed information that could be useful in determining professional development needs during the year was less reliable. Teachers' overall scores were not consistent across the individual observation windows, which limited their reliability as a measure of a teacher's persistent classroom practice and the utility of using ratings from a single observation to identify teachers in need of support. Similarly, scores at the dimension level from a single observation window had insufficient reliability to indicate on which specific dimensions of classroom practice teachers should improve.

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

Chapter 3. Findings About the Intervention's Measure of Student Growth

This chapter presents findings about the intervention's measure of student growth as implemented during the first year of the study. The measure is intended to differentiate teacher performance to identify lower- and higher-performing teachers, and thus identify teachers in need of improvement. In addition, the measure is intended to provide information about a teacher's relative performance in reading/ELA versus mathematics, for those who teach both subjects.

The chapter begins by describing the design of the measure of student growth and discussing findings about how well it was implemented. The chapter then examines how well the measure differentiated teacher performance and describes how well it identified subject areas of relative strength. All findings in this chapter pertain to teachers and principals in the treatment schools.

Key Findings

Implementation

- A large majority (80 percent) of teachers had a sufficient number of students with the achievement data required to estimate value-added scores.
- Most teachers (85 percent) and principals (81 percent) participated in the student growth report training, but less than half (39 percent of teachers with value-added scores, 40 percent of principals) accessed the reports.

Characteristics of the Student Growth Performance Information

- Many teachers with a student growth report (23 percent in reading/ELA, 53 percent in mathematics) had a value-added score that measurably differed from the district average.
- The value-added scores provided some reliable information to distinguish between lower- and higher-performing teachers.
- Among teachers with value-added scores in both reading/ELA and mathematics, about half (48 percent) had student growth reports that suggested the teacher performed better in one subject area than another.

Overview of the Intervention's Measure of Student Growth

The measure of student growth was designed to provide teachers with information about their contribution to their students' achievement growth relative to other teachers in their districts (i.e., value-added scores). AIR estimated individual teachers' value-added scores with a statistical method for analyzing multiple years of students' test score data. (See appendix F for technical details about the estimation.) A teacher's value-added score indicates how much a teacher's students gained, on average, compared to similar students in the district (i.e., those in the same grade, with similar prior performance and other characteristics). A value-added score of zero means that on average the teacher's students performed exactly as expected, or not differently from students with the same prior test scores and characteristics. A positive value-added score indicates that a teacher's students performed, on average, better than they would have performed with an average teacher. For example, a value-added score of 0.2 indicates that a teacher's students had test scores 0.2 standard deviations higher, on average, than if they had an average teacher. A negative value-added score indicates that a teacher's students performed, on average, worse than if they had an average teacher. A negative value-added score does not necessarily mean a teacher is ineffective in an absolute sense, or that his or her students did not make academic gains, because value-added scores depend on the relative performance of other students and teachers in the district.[54]

AIR prepared three waves of student growth reports, as shown in exhibit 3.1. The first wave of reports was released between February and April of the first study year, prior to the study's spring surveys. The second and third waves were released in the fall of the second study year and the fall of the year after the study.

The reports emphasized each teacher's average value-added score over the two previous years because value-added scores can fluctuate significantly from year to year (Goldhaber and Hansen 2013; McCaffrey et al. 2009). Single-year scores were reported for teachers who had value-added scores for only one of the two previous years.

Exhibit 3.1. Timeline for estimating value-added scores and delivering student growth reports

	2010–11	2011–12	2012–13 (Study year 1)	2013–14 (Study year 2)	2014–15
Wave 1	Value-added scores estimated for these years		Delivered spring 2013		
Wave 2		Value-added scores estimated for these years		Delivered fall 2013	
Wave 3			Value-added scores estimated for these years		Delivered fall 2014

Reports were designed to provide information about a teacher's contribution to student achievement overall, and in particular grades and subjects. Each report presented a teacher's overall value-added score, the score for each subject the teacher taught, and the score for each subject-grade combination. To help readers compare a teacher's scores with other teachers' scores, the report presented the percentile rank for the teacher's value-added scores, indicating how well the teacher performed relative to other teachers in the same district. In addition, to help readers draw inferences correctly, the report included information about measurement error, such as the standard errors of the teacher's value-added scores and the confidence intervals of their percentile ranks.[55]

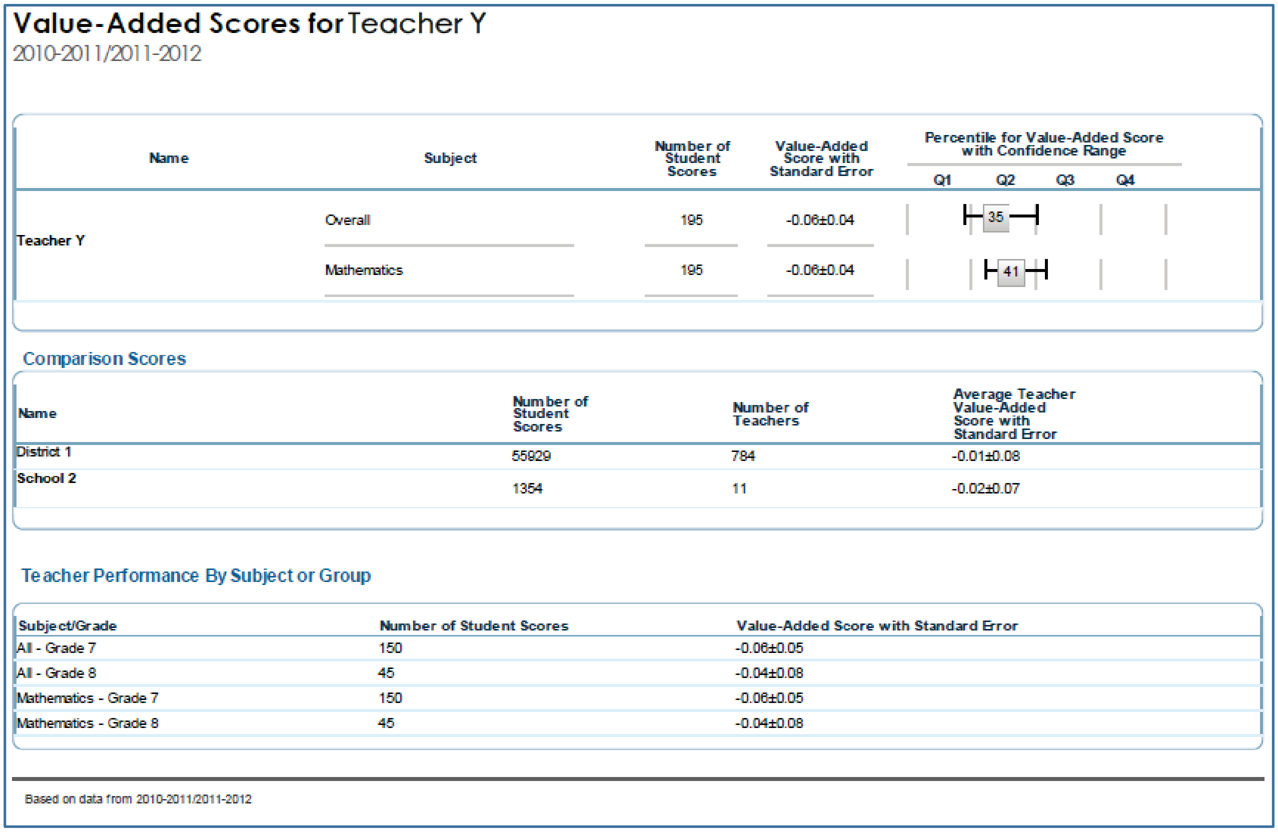
For illustration purposes, exhibit 3.2 shows the main results page from a student growth report for a teacher teaching mathematics in grades 7 and 8 in 2010–11 and 2011–12. The teacher had a

[54] A teacher's value-added score is a measure of a teacher's relative effect on student achievement based on how much students are predicted to learn during the year. Although a value added score is not a direct measure of how much students learned during the year, for readability, we refer to value-added as a measure of student growth.

[55] The student growth reports presented the 80 percent confidence interval for each percentile rank, indicating that there was an 80 percent chance that the interval contained the teacher's true percentile rank.

two-year average value-added score of -0.06 for grade 7 based on 150 student scores and -0.04 for grade 8 based on 45 student scores. The score of -0.06 for grade 7 indicates that the math achievement of the teacher's grade 7 students was 0.06 standard deviations below what would be predicted given the students' background. Since the teacher taught only mathematics, the teacher's overall value-added score across grades and subjects was the same as the teacher's value-added score in mathematics across grades (i.e., -0.06). The teacher ranked at the 35th percentile based on the overall value-added score and at the 41st percentile based on the mathematics value-added score among all teachers with the relevant scores in the district.[56]

Exhibit 3.2. Screenshot of the main results page of a sample student growth report



Value-Added Scores for Teacher Y

2010-2011/2011-2012

Name	Subject	Number of Student Scores	Value-Added Score with Standard Error	Percentile for Value-Added Score with Confidence Range (Q1 Q2 Q3 Q4)
Teacher Y	Overall	195	-0.06±0.04	35
	Mathematics	195	-0.06±0.04	41

Comparison Scores

Name	Number of Student Scores	Number of Teachers	Average Teacher Value-Added Score with Standard Error
District 1	55929	784	-0.01±0.08
School 2	1354	11	-0.02±0.07

Teacher Performance By Subject or Group

Subject/Grade	Number of Student Scores	Value-Added Score with Standard Error
All - Grade 7	150	-0.06±0.05
All - Grade 8	45	-0.04±0.08
Mathematics - Grade 7	150	-0.06±0.05
Mathematics - Grade 8	45	-0.04±0.08

Based on data from 2010-2011/2011-2012

NOTE: The actual names of the teacher, district, and school in the original report are not shown in this exhibit for confidentiality reasons.

SOURCE: AIR value-added system.

AIR also prepared reports for principals on the value-added scores of teachers in their schools. Each report presented a table with the overall value-added score of each teacher in the school, making it possible to compare across teachers. The report also presented individual teachers' value-added scores by subject and grade, as well as the school-average and district-average

[56] The teacher's percentile rank based on the overall value-added score represents the teacher's relative standing among all teachers with an overall value-added score in the district, whereas the teacher's percentile rank based on the mathematics value-added score represents the teacher's relative standing among all teachers with a mathematics value-added score in the district (who were a subset of teachers with an overall value-added score).

value-added scores, overall and by subject and grade. (See appendix J for a sample student growth report for principals.)

Findings About the Implementation of the Intervention's Measure of Student Growth

As indicated in chapter 2, to encourage implementation of the evaluation system performance measures tested in this study, an AIR team separate from the evaluation team monitored implementation and provided support when needed to keep the activities on track. For example, to support the successful dissemination of the student growth reports, the AIR team sent reminder notices to teachers who had not yet accessed their student growth reports. To examine the extent to which the intervention's student growth measure was implemented as intended, we examined how many teachers received value-added scores, whether teachers and principals participated in training related to the student growth reports, and whether teachers and principals accessed the reports. Findings from these analyses are presented next.

A large majority of teachers had a sufficient number of students with the achievement data required to estimate value-added scores. Overall, student achievement data were sufficient to estimate value-added scores for 80 percent of teachers, 68 percent based on two-year averages and an additional 12 percent based on single-year scores. There were not sufficient data to estimate value-added scores for 20 percent of the teachers.

Most teachers and principals participated in the student growth report training. Prior to the release of the student growth reports, the study's implementation team held live training webinars to help teachers and principals understand the meaning of value-added scores, the content of the reports, and how to access the reports. Overall, 85 percent of teachers and 81 percent of principals participated in the webinars.

Less than half of teachers and principals accessed the reports, with access rates varying substantially across schools. Despite good attendance at the webinars, access rates were low—39 percent of the teachers with value-added scores and 40 percent of the principals accessed the reports.[57] Teacher access rates varied widely across schools. In nearly a quarter (23 percent) of the schools, none of the teachers in the relevant grades and subjects accessed their student growth reports; in contrast, in 15 percent of the schools, all teachers accessed their reports. The access rates also varied substantially across districts among both teachers and principals (see exhibit G.1 in appendix G).

Teachers in schools where the principal accessed the student growth reports were almost three times more likely to access their student growth report than teachers in schools where the principal did not access the student growth reports (odds ratio = 2.93; p=.027).[58]

[57] 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 enough data to estimate value-added scores. This included all but one school in the sample.

[58] To examine how teacher access rates were associated with whether the principal accessed the student growth reports, we used a two-level logistic regression model (teachers nested in schools) that predicted whether teachers

Findings About the Student Growth Information

The student growth reports prepared for treatment teachers and principals were intended to provide information that would differentiate teachers based on their contribution to student growth. Given that purpose, a key question is whether the value-added scores are reliable. One way to assess the reliability of value-added scores is to examine the degree to which a teacher's value-added score is stable from one year to the next. In this section, we report on the potential utility of the student growth reports by first describing the scores provided in the student growth reports and then examining how well the scores differentiated teachers and provided them with information about their relative performance in reading/ELA versus mathematics.

Treatment teachers' value-added scores were distributed as expected. Consistent with the district-wide distributions, about half of the treatment teachers (48 percent) had overall value-added scores below the district average, and about half (52 percent) had overall value-added scores above the district average. Mathematics value-added scores were more spread out than reading/ELA value-added scores (see exhibit 3.3), which indicates that differences between low and high value-added teachers were more pronounced in mathematics than in reading/ELA.[59] For example, 8 percent of teachers had a mathematics value-added score below -0.20 standard deviations (in student test score units), and 13 percent had a mathematics value-added score above 0.20 standard deviations. By comparison, 3 percent of teachers had a reading/ELA value-added score below -0.20 standard deviations (in student test score units), and 2 percent had a reading/ELA value-added score above 0.20 standard deviations. The means and standard deviations of the value-added scores, by teacher characteristics, are presented in exhibit G.2 in appendix G.

accessed their student growth report based on the school district and whether the principal accessed the student growth reports.

[59] The finding that teachers varied more in their mathematics value-added scores than in their reading/ELA scores is consistent with findings from other studies of value-added scores (Hanushek and Rivkin 2010; Taylor and Tyler 2012).

Exhibit 3.3. Distribution of treatment teachers based on their value-added scores

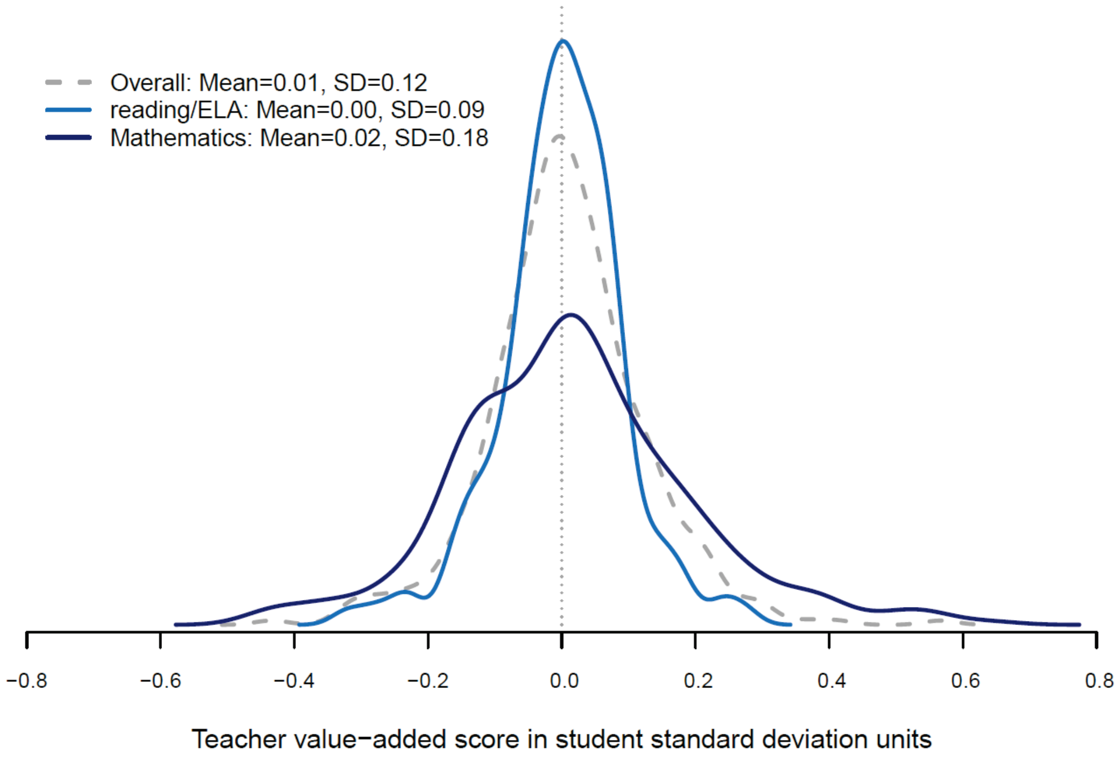


Exhibit Reads: The overall value-added scores were concentrated around zero, with few scores below -0.20 or above 0.20.

NOTE: The exhibit shows the density of teachers across the value-added score distribution, where the area under each curve between two scores represents the percentage of teachers with scores in that range, and the total area under the curve sums to 100 percent. Value-added scores are in student test score standard deviation units. Sample size = 433 teachers with overall value-added scores, 326 teachers with reading/ELA value-added scores, and 342 teachers with mathematics value-added scores.

SOURCE: AIR value-added system.

Many teachers with a student growth report had a value-added score that measurably differed from the district average, particularly in mathematics. As with all value-added measures, uncertainty in a teacher's value-added score means teachers may not truly differ in performance from one another even if their estimated scores are different. To indicate the amount of uncertainty around each teacher's score, the student growth reports included 80 percent confidence intervals, which showed the range of scores that have an 80 percent chance of including the teacher's "true" score. This benchmark was selected in order to appropriately balance two types of risks within the context of an intervention designed to provide feedback on performance without explicit consequences such as promotion or dismissal: (1) the risk of misidentifying truly average teachers as below- or above-average, and (2) the risk of misidentifying teachers who were truly below- or above-average as average teachers.[60] Taking

[60] The two types of risk reflect Type I and Type II errors, respectively. As the confidence interval becomes wider, the Type I error rate decreases, but the Type II error rate increases. See Schochet and Chiang (2013) for an analysis of the magnitude of Type I and Type II errors if teachers are identified as average versus above or below average, based on a value-added model similar to the model used in this study. The results indicate that with two years of value-added data, the Type I error must be set at about 20 percent (corresponding to an 80 percent confidence interval) to achieve a Type II error of similar size (20 percent), under reasonable assumptions. Similarly,

into account the confidence interval for each teacher's value-added scores, some teachers could infer that they improved student achievement "measurably" more than, or less than, a teacher with the district average score (see exhibit 3.4).[61] Based on the reading/ELA value-added scores, 12 percent of teachers had a value-added score that was considered measurably above the district average and 11 percent had a score considered measurably below average. Based on the mathematics value-added scores, 28 percent of teachers had a value-added score that was considered measurably above the district average, and 25 percent had a score considered measurably below average.[62]

Raudenbush and Jean (2012) discuss the tradeoff between a 95 and 75 percent confidence interval, noting that teachers might wish to use the latter for self-evaluation.

[61] A teacher's value-added score was considered measurably different from the district average if the score's 80 percent confidence interval (which was the confidence interval used in the student growth reports) did not include the district average score.

[62] If a 95 percent confidence interval is used to determine whether teachers are measurably different from average instead of the 80 percent confidence interval used for the student growth reports, fewer treatment teachers would be considered measurably above/below average. For Reading/English language arts, 91 percent would not be measurably different from average, 4 percent would be measurably above average, and 5 percent would be measurably below average. For mathematics, 67 percent would not be measurably different from average, 18 percent would be measurably above average, and 15 percent would be measurably below average.

Exhibit 3.4. Distribution of treatment teachers based on whether their value-added score was considered measurably above or below the district average, overall and by subject

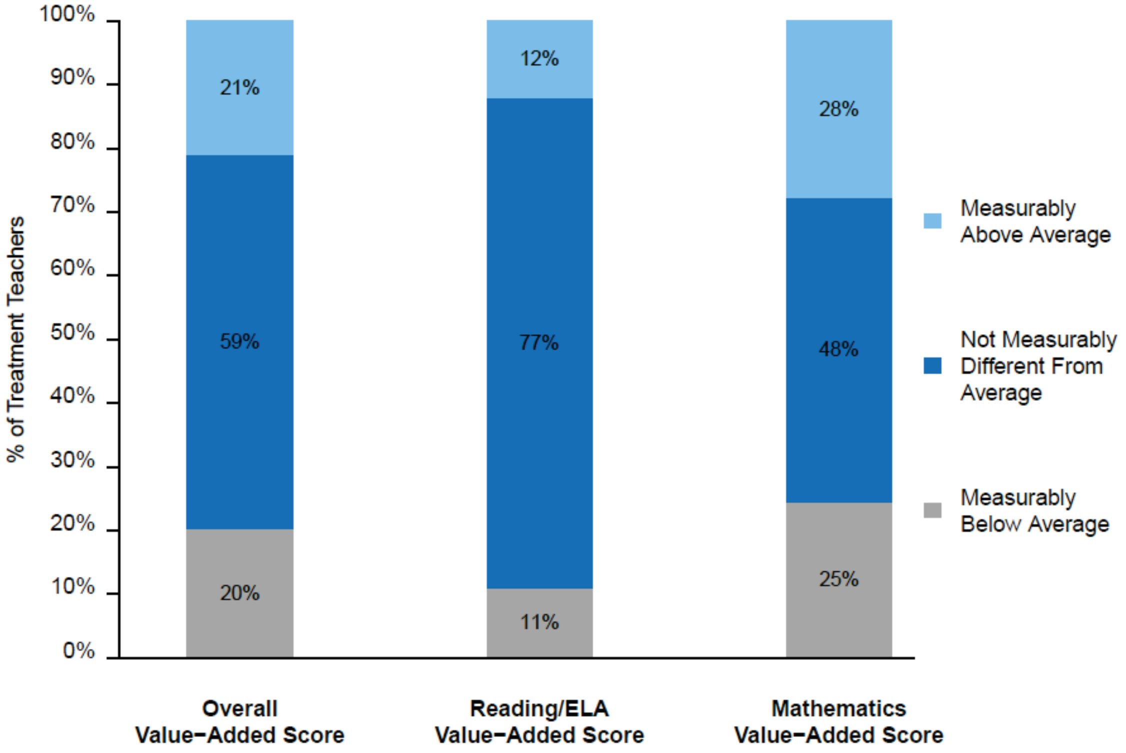


Exhibit Reads: For treatment teachers with reading/ELA value-added scores, 12 percent had scores considered measurably above the district average.

NOTE*:* The distributions of teachers are based on whether the 80 percent confidence interval for a teacher's value-added score was above or below the district average. Sample size = 433 teachers with overall value-added scores; 326 teachers with reading/ELA value-added scores; and 342 teachers with mathematics value-added scores. Reported percentages may not sum to 100 percent because of rounding.

SOURCE: AIR value-added system.

The value-added scores provided some reliable information to distinguish between lower- and higher-performing teachers. To distinguish between lower- and higher-performing teachers, and thus help identify teachers in need of improvement, the value-added scores need to be sufficiently reliable to identify lower-performing and higher-performing teachers. This means a teacher's value-added score should reflect persistent performance and not mainly idiosyncratic factors introduced by the classroom composition or abnormal events. We estimated the degree to which the value-added scores were a reliable measure of a teacher's persistent performance based on how much a teacher's value-added score varied across the two years of student growth data that were used to estimate the value-added scores. (See appendix C for details about the methods and results.) These reliability estimates tell us how consistent, or stable, the value-added scores were over two years of classroom instruction. Based on two years of student growth data, the value-added score reliability was estimated to be .44 for reading/ELA and .68 for mathematics. These reliability estimates are consistent with estimates found in research on other value-added measures, which range from about .20 to .85 depending on factors such as the number of students included in the value-added model, the grade level, subject, and type of assessment used to measure student achievement (Goldhaber and Hansen 2013; McCaffrey et al. 2009; Mihaly et al. 2013; Whitehurst, Chingos and Lindquist 2014).

Among teachers with value-added scores in both reading/English language arts (ELA) and mathematics, about half had student growth reports that suggested the teacher performed better in one subject area than the other. Of the 433 teachers with value-added scores, 55 percent had value-added scores for both reading/ELA and mathematics (e.g., teachers in self-contained elementary school classrooms). By comparing their performance categories in reading/ELA and mathematics, teachers could draw conclusions about whether their performance differed in the two subjects. In particular, based on the 80 percent confidence intervals used for the student growth reports, teachers could infer whether their performance in each subject was measurably below average, not measurably different from average, or measurably above average. In total, 48 percent of teachers with scores in both subjects had student growth reports that suggested different performance in reading/ELA than mathematics.[63] In particular, 21 percent of the teachers had student growth reports that suggested the teacher performed better in reading/ELA than mathematics, and 26 percent had student growth reports that suggested the teacher performed better in mathematics than reading/ELA. (See exhibit G.3 in appendix G.)[64,65]

Summary

Study districts were successful at implementing the study's student growth measure in some respects. On the one hand, they successfully supplied the data needed, and value-added scores were computed for a large majority of teachers. In addition, a majority of the student growth reports were based on two years of value-added data, as intended. On the other hand, although most teachers and principals attended training webinars prior to the release of the reports, fewer than half of teachers and principals accessed their reports. Meanwhile, the value-added measure distinguished teacher performance but only to a degree. The scores varied, as expected, and were reliable enough to identify some teachers who contributed more to student growth than a teacher with the district average value-added score, and some teachers who contributed less. Among teachers with scores in both reading/ELA and mathematics, about half had student growth reports that suggested that the teacher performed better in one subject than the other.

[63] We examined differences in a teacher's subject-specific value-added scores, which are based on student growth in test score standard deviation units in each subject. The student growth reports also included the teacher's value-added percentile ranking in each subject. We based the analysis on the test score standard deviation units, rather than the percentile rankings, because the test score metric is used to estimate each teacher's value-added scores, and it is the metric used to report value-added scores in this chapter.

[64] The total percentage of teachers who had student growth reports that suggested different performance in reading/ELA than mathematics (48 percent) is not equal to the sum of the percentage that performed better in reading/ELA (21 percent) plus the percentage that performed better in mathematics (26 percent) due to rounding.

[65] We also estimated the degree to which the difference between a teacher's value-added scores in reading/ELA and mathematics is a reliable measure of the teacher's true relative performance in the two subjects. The estimated reliability of the difference between a teacher's subject-specific value-added scores was .52. (See appendix C for details about the estimation method and results.)

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

Chapter 4. Findings About the Intervention's Measure of Principal Leadership

This chapter presents findings about the intervention's measure of principal leadership as implemented during the first year of the study. The information provided as feedback by the measure, the Vanderbilt Assessment of Leadership in Education (VAL-ED), is intended to. differentiate principal performance, to identify lower- and higher-performing principals, potentially suggesting principals for additional support. In addition, the measure is intended to identify practices that, if improved, would lead to more effective leadership and higher student achievement.

The chapter begins by describing the design of the principal leadership measure and feedback. It then discusses findings about how well the measure and feedback were implemented, examines how well the measure differentiated principal performance, and describes how well it identified practices to improve. All findings in this chapter pertain to principals in the treatment schools.

Key Findings

Implementation

- All principals and their supervisors received training on using VAL-ED.
- All VAL-ED reports incorporated input from the principal, the principal's supervisor, and most teachers (80 to 90 percent), as intended.
- All VAL-ED feedback sessions occurred as planned.

Characteristics of the Principal Leadership Information

- The VAL-ED ratings classified some principals as lower-performing (14 to 27 percent) and some as higher-performing (8 percent), as intended.
- VAL-ED ratings provided by principals, supervisors, and teachers in the fall were often too different to form a reliable measure, but the spring ratings were consistent enough to distinguish between some lower- and higher-performing principals and were positively correlated with another survey measure of principal instructional leadership.
- Nearly all principals (more than 95 percent) received VAL-ED scores that differed across different dimensions of principal leadership, but the scores did not reliably distinguish between the dimensions.

Overview of the Intervention's Measure of Principal Leadership

This intervention's measure of principal leadership was the VAL-ED, a 360-degree survey of principals, their supervisors, and teachers. In the study's intervention, the VAL-ED was used to provide principals and their supervisors with information about principal leadership, and was administered in both fall and spring. The VAL-ED is designed to measure leadership behaviors associated with student learning. The dimensions measured include six "core components" and six "key processes," as listed in exhibit 4.1 (see exhibit H.1 in appendix H for definitions of the core components and key processes). In addition, it measures leadership in each of the 36 "component-by-process" performance areas. For example, one of the performance areas pertains to how effective the principal was in developing plans for setting high standards for student learning, which is the intersection of the key process "Planning" and the core component "High standards for student learning."

Exhibit 4.1. VAL-ED core components and key processes

Core components	Key processes
• High standards for student learning	• Planning
• Rigorous curriculum	• Implementing
• Quality instruction	• Supporting
• Culture of learning and professional behavior	• Advocating
• Connections to external communities	• Communicating
• Systemic performance accountability	• Monitoring

To obtain the performance information, the VAL-ED survey asks each respondent to use a 5-point scale (1 = *ineffective*; 2 = *minimally effective*, 3 = *satisfactorily effective*, 4 = *highly effective*, and 5 = *outstandingly effective*) to rate a principal's effectiveness in 72 leadership behaviors that represent the 36 component-by-process areas.[66] (See exhibit H.2 in appendix H for a sample of survey items.) The online system collects the responses electronically and produces a report on the principal.

The report is the focus of the feedback session between the principal and his or her supervisor. The VAL-ED training for principal supervisors is designed to prepare him or her to provide principals with structured feedback. Specifically, the feedback sessions were expected to cover: definitions of the core components and key processes; the overall results; the results received from each of the three respondent groups (i.e., teachers, principals, and principal supervisors); and identification of dimensions on which the principal is strong and dimensions on which the principal plans to grow.

The report is generated by the online system automatically, based on survey responses only. It presents an overall score, a score for each core component, and a score for each key process based on the average responses across the three respondent groups (i.e., principal, supervisor, and teachers), with each group weighted equally. The report also presents the percentile ranks corresponding to the principal's overall score, core component scores, and key process scores based on how the principal performed relative to the principals included in a national VAL-ED field test. To aid principals and their supervisors in interpreting the ratings, the developer assigned each score a performance level (*below basic*, *basic*, *proficient*, or *distinguished*).[67] (See exhibit H.3 in appendix H for a screenshot from a sample VAL-ED report that includes the performance level descriptors.) Scores are also reported separately by respondent group. (See exhibit H.4 in appendix H.) The process used to translate VAL-ED scores to percentile ranks and

[66] In both fall and spring, each principal and the principal supervisor took the full 72-item survey, and each teacher took a 36-item survey with one item for each of the 36 component-by-process areas.

[67] The developer used a standards-setting process and national field test data to set the performance level cut scores (Porter et al. 2008). The range of scores corresponding to each performance level is as follows: 1.00–3.28: *below basic*, 3.29–3.59: *basic*, 3.60–3.99: *proficient*, and 4.00–5.00: *distinguished*. The cut scores resulted in the following distribution of principals in the national field test data: 17 percent at the *below basic* level, 33 percent at the *basic* level, 36 percent at the *proficient* level, and 14 percent at the *distinguished* level (Porter et al. 2010).

performance levels does not adjust for the point during the school year when the VAL-ED survey was administered (i.e., fall versus spring).

In addition to the overall score and scores for core components and key processes, the VAL-ED report presents the score for each component-by-process combination in a six-by-six matrix, with color-coded cells indicating performance level. (See exhibit H.5 in appendix H.) The report concludes with a list of leadership behaviors in up to six lowest-rated component-by-process areas, which the report labels "leadership behaviors for possible improvement."

Findings About the Implementation of the Intervention's Measure of Principal Leadership

As mentioned in chapter 2, to encourage implementation with fidelity, an AIR team separate from the evaluation team monitored implementation and provided support when needed to keep the activities on track. For example, AIR staff monitored the percentage of teachers who provided input on the principal's leadership (i.e., the VAL-ED survey response rate) and offered assistance if the response rates were not on track to reach 80 percent within each school. This section presents findings about the extent to which the measure was implemented as intended, focusing on participation in VAL-ED training and feedback sessions and VAL-ED survey response rates.

All principals and their supervisors received training on using VAL-ED*.* All principals and their supervisors participated in a two-hour training in summer 2012. During the school year, just prior to the feedback sessions in the fall as well as the spring, all principal supervisors also attended a one-hour training designed to prepare them to conduct the feedback sessions. In addition, teachers were offered a webinar during the school year that was designed to prepare them to complete the VAL-ED survey.[68]

All VAL-ED reports incorporated input from the principal, the principal's supervisor, and most teachers. All principals and their supervisors took the VAL-ED surveys in both fall and spring. The majority of teachers in each school (80 percent in fall and 90 percent in spring, on average) also completed the survey.

All VAL-ED feedback sessions occurred as planned. In both fall and spring, all principals met with their supervisors to discuss their VAL-ED reports. The supervisors reported that the feedback sessions on average lasted 52 minutes in fall and 46 minutes in spring.

Findings About the Principal Leadership Information Principals Received

As described earlier in the chapter, the principal leadership measure provided detailed information for principals on their leadership. The information was conveyed after each VAL-ED survey administration, both through a formal report, generated by the online system, and

[68] AIR did not track attendance at the respondent webinar. The vendor warned that doing so could discourage teachers from completing the VAL-ED survey by making them feel that their participation in the VAL-ED survey was not anonymous.

through a feedback session with the principal's supervisor. The VAL-ED reports included overall mean scores, ranging from 1 to 5, as well as mean scores for each of six core components and six key processes. They also provided mean scores for each of the 36 component-by-process performance areas. The reports translated each of these mean scores into a performance level (i.e., *below basic*, *basic*, *proficient*, or *distinguished*) and a percentile rank. Finally, each of the scores was additionally reported by respondent group (i.e., separate principal, supervisor, and teacher scores), showing the extent to which the respondent groups agreed with each other.

The study data do not allow us to determine specifically what information, if any, principals focused on. For analyses in the first part below, we look at overall ratings within each assessment window (i.e., fall and spring). For analyses in the second part, we look at dimension ratings (i.e., core component scores and key process ratings). For both overall and dimension ratings, we examine the variation in the scores and whether the measure was sufficiently reliable to distinguish performance among principals in this study.

Variation in Principal Performance Based on Overall Ratings

Although nearly all of the survey respondents rated principals as "satisfactorily effective" (the midpoint on the rating scale) or higher, the VAL-ED reports converted those scores into the full range of performance levels, labeling many principals "below basic." The VAL-ED reports showed each score and the corresponding performance level (i.e., *below basic*, *basic*, *proficient*, or *distinguished*) to communicate the general quality of principal leadership. Nearly all principals (92 percent in fall and 97 percent in spring) had an overall score of 3.0 or higher, where 3 is defined on the survey as *satisfactorily effective*.[69] However, the cut scores used to convert a principal's scores into performance levels placed many principals with scores in the upper-half of the rating scale (i.e., 3 or above) in the bottom two performance levels (*below basic* or *basic*). In both fall and spring, principals were distributed across all four performance levels (see exhibit 4.2). Based on their ratings in the fall, 70 percent of principals were in the bottom two performance levels (i.e., *below basic* and *basic*). In the spring, about half of the principals (51 percent) were in the bottom two performance levels (see exhibit 4.3).

[69] Descriptive statistics for the scores are presented in exhibits H.6 and H.7 in appendix H.

Exhibit 4.2. Distribution of treatment principals based on their VAL-ED overall scores in fall and spring

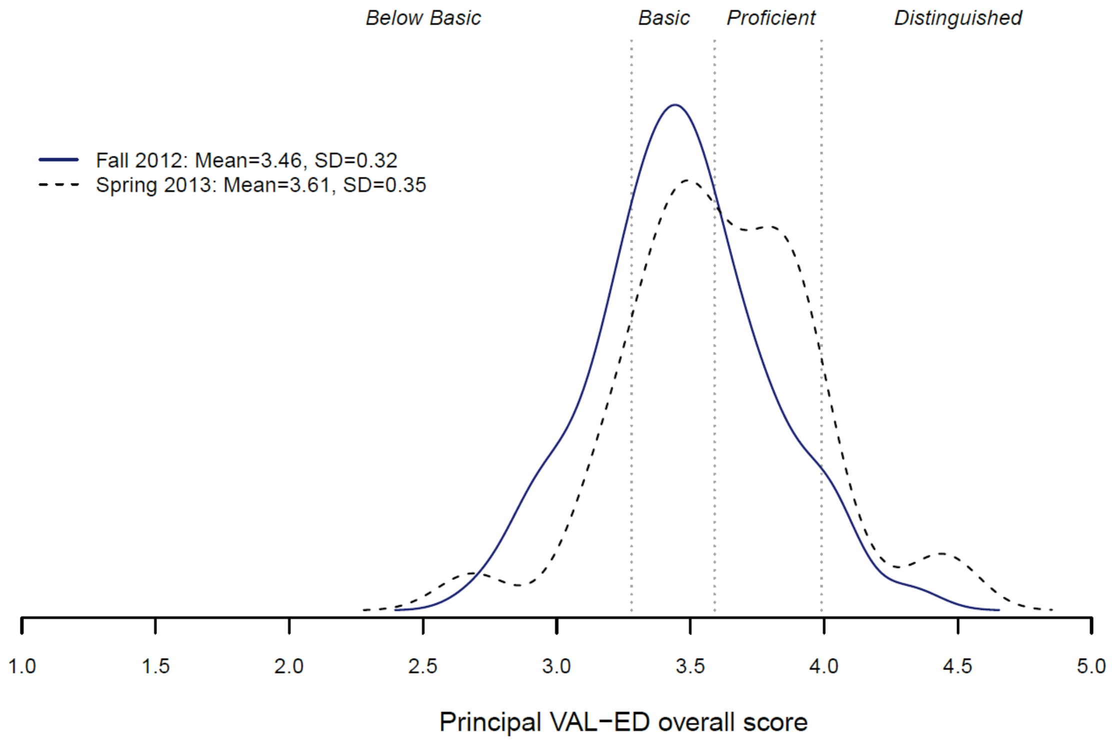


Exhibit Reads: Treatment principals were distributed across the four performance levels based on their fall 2012 VAL-ED overall scores, with a score just below 3.5 being the most common overall score.

NOTE: Sample size = 63 principals for both fall 2012 and spring 2013.

SOURCE: Fall 2012 and Spring 2013 VAL-ED Surveys.

Exhibit 4.3. Distribution of treatment principals across performance levels based on VAL-ED overall scores in fall and spring

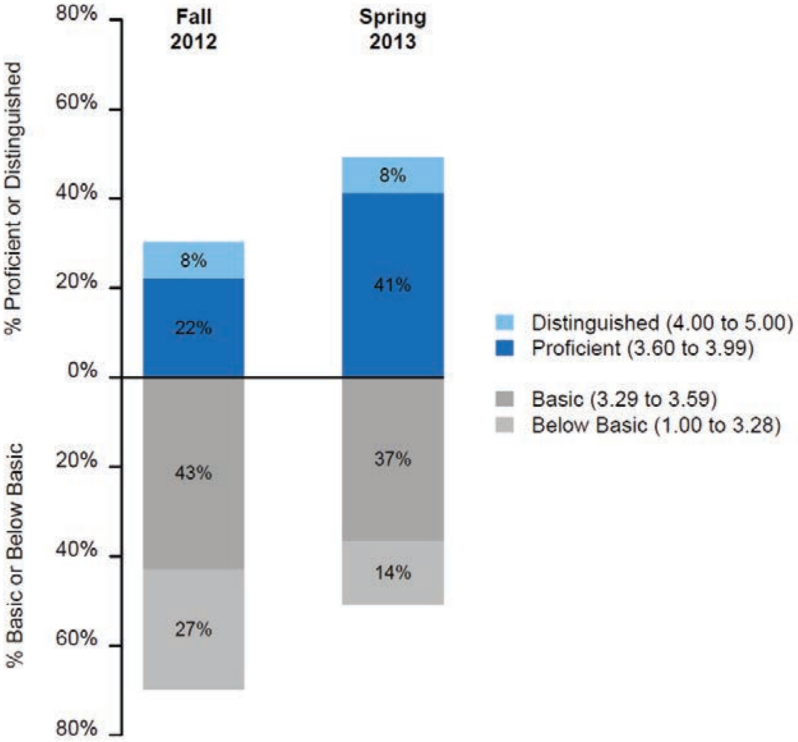


Exhibit Reads: In fall 2012, 8 percent of treatment principals had a VAL-ED overall score at the *distinguished* level, 22 percent at the *proficient* level, 43 percent at the *basic* level, and 27 percent at the *below basic* level.

NOTE: Performance level distributions are based on principals' VAL-ED overall scores at each assessment window. The overall score is an average of the scores from the principal's supervisor, teachers, and the principal's own self-rated score, with each group weighted equally. Sample size = 63 principals for both fall 2012 and spring 2013. Reported percentages may not sum to 100 percent because of rounding. Sample size = 63 principals for both fall 2012 and spring 2013.

SOURCE: Fall 2012 and Spring 2013 VAL-ED Surveys.

In addition to performance levels, each principal received a percentile ranking indicating how the principal's overall score ranked relative to a national sample of principals. Half of the principals had an overall score at or below the 29th percentile in the fall and at or below the 46th percentile in the spring. The increase in average VAL-ED overall scores from the fall to spring is primarily a product of an increase in the principal self-ratings, which is described next.

Average ratings of principal leadership from the three respondent groups were similar in the fall; however, in the spring, principal self-ratings were higher on average than the ratings from their supervisor and teachers. In addition to the overall score averaged across the three respondent groups, the VAL-ED reports included information about how each respondent group rated the principal, which allowed the principal and supervisor to see whether the three respondent groups had similar opinions of the principal's effectiveness. With ratings from the different respondent groups, one concern is whether the respondent groups provide different ratings, on average. For example, do principals tend to give themselves more favorable ratings than the teachers and supervisors, or do the teachers tend to provide lower ratings than the principal or supervisor? In the fall, the average overall scores were similar across the respondent groups, though the shape of the distributions for supervisor and teacher ratings differed significantly (see exhibit 4.4).[70] From fall to spring, the average overall score from the principal self-ratings increased significantly from 3.43 to 3.76, which explains most of the increase in overall scores from fall to spring. Thus, by spring, principals had become the most lenient of the three rater groups: In the spring, the average principal self-rating was significantly higher than the average rating given by supervisors (3.50) and teachers (3.57).

[70] Statistically significant differences in distributions were assessed with Kolmogorov-Smirnov two-sample tests ($p < .05$). In addition to the fall supervisor and teacher ratings distributions, there were statistically significant differences in the supervisor and principal ratings in the spring. The distribution of principal ratings also differed in the fall and spring.

Exhibit 4.4. Distribution of treatment principals based on their VAL-ED overall scores in fall and spring, by respondent group

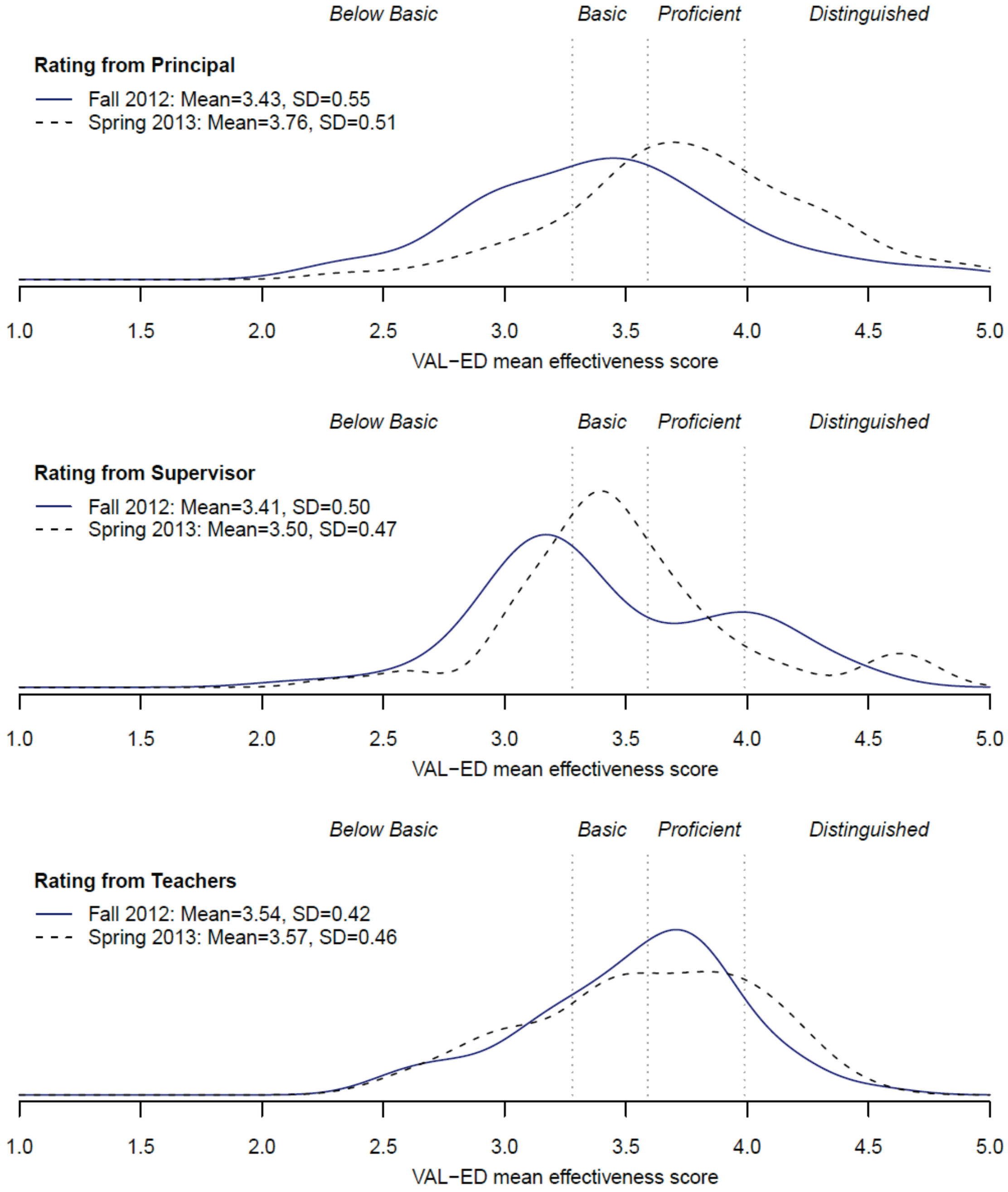


Exhibit Reads: Treatment principals were distributed across the four performance levels based on their fall 2012 VAL-ED self-rated respondent group overall score, with principals and supervisors giving more mean scores below 3.5 than teachers.

NOTE: Sample size = 63 principals for both fall 2012 and spring 2013.

SOURCE: Fall 2012 and Spring 2013 VAL-ED Surveys.

VAL-ED ratings provided by principals, supervisors, and teachers in the fall were often too different to form a reliable measure, but the spring ratings were consistent enough to distinguish between some lower- and higher-performing principals. To provide information about a principal's overall effectiveness, the VAL-ED scores from each of the three respondent groups should communicate a consistent (i.e., reliable) message about the principal's effectiveness. Based on the literature on 360-degree surveys, we would expect correlations between respondent group scores between .25 and .35.[71] In the fall, however, agreement among the three respondent groups' overall scores was low, with some correlations below .10 (see exhibit 4.5). Principal-supervisor agreement and principal-teacher agreement were particularly low in the fall, with correlations less than .10. In the spring, correlations were higher, and thus the reports provided principals and supervisors a more consistent message about a principal's effectiveness. We estimated that the VAL-ED overall score reliability (i.e., inter-rater reliability) was .19 in the fall and .51 in the spring (see appendix C for details about the estimation methods and results).[72] The improved reliability in spring reflects greater agreement between the principal self-ratings and the other two respondent groups.[73]

Exhibit 4.5. Correlations between VAL-ED respondent group overall scores from different respondent groups in fall and spring

Correlation	Fall 2012	Spring 2013
Principal and supervisor	.08	.27*
Principal and teachers	.06	.26*
Supervisor and teachers	.27*	.38*

Exhibit Reads: The correlation between VAL-ED respondent group overall scores from principal self-ratings and supervisor ratings was .08 in the fall.

NOTE: Sample size = 63 principals for both fall 2012 and spring 2013.

* Significantly different from zero with *p* < .05.

SOURCE: Fall 2012 and Spring 2013 VAL-ED Surveys.

The increase between fall and spring in the correlations between principal, supervisor, and teacher ratings may appear to contradict a finding reported earlier -- that average ratings provided by the three respondent groups were similar in the fall, but in the spring, principal self-ratings were higher on average than the ratings given by their supervisor and teachers. However, they are not inconsistent. The earlier conclusion concerns the *average* rating given by principals,

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

[72] As a point of reference, reliability for the classroom observation four-window average scores was estimated to be between .42 and .75.

[73] A principal's VAL-ED score for the teacher respondent group is based on the average score from all teachers that filled out the VAL-ED survey about the principal. Since multiple teachers in a school rated the principal, we can estimate the extent to which teachers in a school gave the principal similar overall VAL-ED scores. For the fall, 76 percent of the variation in teacher ratings was within principal, and the other 24 percent was between principal, implying an inter-rater reliability of .24. For the spring, the inter-rater reliability was .25. The overall reliability of the teachers' rating of their principal depends on the number of teachers that rated the principal. On average, about 30 teachers rated a principal, which implies the teacher score had, on average, reliability of .91 in both the fall and spring.

supervisors, and teachers, while the one here concerns the *correlation* among ratings – for example, the degree to which principals who gave themselves relatively high ratings also received relatively high ratings from their supervisor and teachers.

VAL-ED ratings were positively correlated with another survey measure of principal instructional leadership, providing some evidence for the validity of the VAL-ED ratings. As a check on the validity of the VAL-ED scores, we examined the correlation of principals' overall VAL-ED scores in the spring with a measure of principal instructional leadership that was based on the study's spring 2013 teacher survey.[74] Both the overall VAL-ED score and overall scores for each respondent group had a positive relationship with the teacher survey principal instructional leadership measure. The correlation between the spring VAL-ED overall score and the principal leadership measure was .56.[75]

Feedback on Specific Dimensions of Principal Leadership

To inform decisions about improving practice and identifying professional development needs, the VAL-ED reports provide performance information on different dimensions of leadership (i.e., the six core components and six key processes, and the 36 intersections of components and processes). We begin this section by looking at the variation in scores across the 36 dimensions of principal leadership, which were presented in the VAL-ED reports to identify leadership behaviors for possible improvement. The remainder of the section presents findings on the extent to which the VAL-ED reports provided teachers with reliable information for the core components and key processes, which are the different leadership behaviors measured by the VAL-ED and are used to define the 36 dimensions of principal leadership.

Nearly all principals received VAL-ED scores that differed across different dimensions of principal leadership. To inform decisions regarding professional development, the VAL-ED reports included scores and performance levels for 36 dimensions of principal leadership based on the intersection of the VAL-ED's six core components and six key processes. If a principal received different ratings on different dimensions of their leadership, then that might allow the principal to draw conclusions about dimensions of leadership on which he or she performed relatively well or relatively poorly. An analysis of the extent to which principals' scores on the 36 dimensions of leadership spanned multiple performance levels indicates that in both fall and spring, nearly all principals (more than 95 percent) received scores

[74] The principal instructional leadership measure from the spring 2013 teacher survey is based on the school average Rasch scale of eight teacher survey items adapted from the Chicago Consortium of School Research teacher survey (Chicago Consortium on School Research 2012). There is some evidence that the principal instructional leadership measure is positively associated with classroom instruction and student achievement (Sebastian and Allensworth 2012). Therefore, a positive correlation between the VAL-ED spring scores and the spring 2013 teacher survey provides some evidence of convergent validity.

[75] The spring VAL-ED overall score based on the teacher respondent group had the strongest association with principal leadership (correlation = .90), which is as expected because most of the teachers who rated their principal for VAL-ED were the same teachers who completed the survey that included the principal instructional leadership measure. Correlations of the spring VAL-ED overall scores based on supervisors and on the principal self-ratings were quite a bit smaller (.25 and .12, respectively) but still provide some evidence for the validity of the VAL-ED overall scores.

at multiple performance levels (see exhibit 4.6), with 84 percent of principals in the fall and 89 percent in the spring receiving scores at three or four different performance levels.

Exhibit 4.6. Percentage of treatment principals whose VAL-ED scores spanned one, two, three, or four performance levels in fall and spring

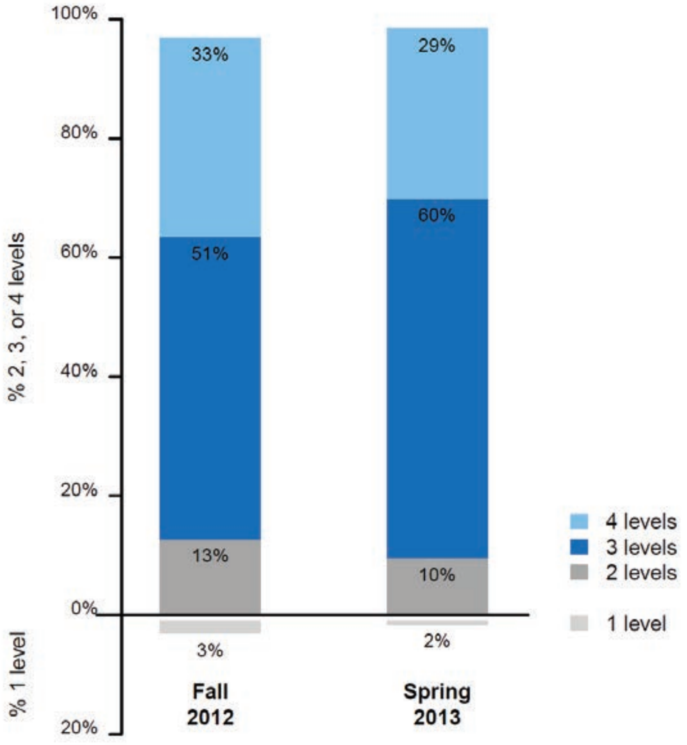


Exhibit Reads: In fall 2012, 33 percent of treatment principals had scores that fell into four different performance levels, 51 percent had scores in three different performance levels, 13 percent had scores in two different performance levels, and less than 5 percent had all scores in the same performance level.

NOTE: Performance level counts are based on the 36 core-component-by-key-process scores for each principal at each assessment window. Reported percentages may not sum to 100 percent because of rounding.

Sample size = 63 principals for both fall 2012 and spring 2013.

SOURCE: Fall 2012 and Spring 2013 VAL-ED Surveys.

Although most principals received scores that spanned multiple performance levels, their scores did not reliably distinguish between different dimensions of their leadership. Even though most principals received scores that spanned multiple performance levels, a principal’s scores may not have clearly distinguished between the dimensions of his or her performance if the scores from different respondent groups did not convey a consistent message about the principal’s relative performance across dimensions of leadership. One way to examine whether scores from the three respondent groups provided principals with a consistent message about the dimension of leadership is to look at whether respondent groups agreed on the principal’s lowest scoring dimension of leadership. Less than a third of the principals had the same lowest-scored dimension of leadership from each of the three respondent groups (25 percent in fall and 30 percent in spring for the core components; 10 percent in fall and 14 percent in spring for the key processes).

Another way to examine the consistency in a principal's dimension scores across respondent groups is to examine the degree to which a principal's scores from the three respondent groups were a reliable measure of whether a principal's performance was better in some dimensions than others. We conducted separate analyses of reliability for the core components and key processes (see appendix C for details about the estimation methods and results). We estimated that the reliability of the difference between a principal's scores for two different core components, on average, was .36 in the fall and .50 in the spring. The reliability of the difference between a principal's scores for two different key processes, on average, was .29 in the fall and .20 in the spring. Thus, within the same report, a principal typically received a different message from the three respondent groups about the dimensions a principal needed to improve on the most. (Correlations among the VAL-ED scores for different dimensions of leadership are provided in exhibits H.8 and H.9 in appendix H.)

Summary

Study districts were successful in implementing the principal leadership performance measure. Principals and their supervisors were trained for their roles, and all of the planned feedback sessions occurred. For the fall and spring assessment windows, all VAL-ED reports incorporated input from the principal, the principal's supervisor, and most teachers, and VAL-ED scores correlated with another measure of leadership included in the study's teacher survey. The VAL-ED also provided performance information that categorized principals as lower- or higher-performing. However, differences in ratings across survey respondent groups (i.e., principal, principal's supervisor, and teachers) limited the consistency of the performance information provided. Likewise, although most of the VAL-ED reports identified at least one dimension for improvement, the respondent groups did not typically agree on the dimensions needing improvement.

Chapter 5. Findings About Educators' Performance Evaluation Experiences

The study's intervention was intended to provide educators with performance information that was more frequent, systematic, and useful as a guide for professional growth than the information that they normally receive. To assess whether this occurred, we analyzed survey data to examine the differences between treatment and control schools in educators' experiences with performance feedback.

The analyses that involve teachers focus on grades 4-8, which were the main grades studied.[76] Separate findings for CLASS districts and FFT districts as well as findings for K–3 teachers are presented in appendix I. Differences in results between the CLASS and FFT districts should be interpreted with caution.   The CLASS and FFT instruments were not randomly assigned to districts. Therefore, any differences in results between CLASS and FFT districts cannot necessarily be attributed to the CLASS and FFT instruments; they may be due to other district characteristics. Unless otherwise noted, all differences discussed in this chapter are statistically significant at the .05 level based on two-tailed tests.

Key Findings

Teachers' Experiences

- Treatment teachers reported receiving more than four times as many feedback sessions with ratings and a written narrative on their classroom practice as control teachers (3.0 versus 0.7 sessions), and they were more likely to receive feedback on their value-added than were control teachers (45 versus 24 percent).
- Among treatment and control teachers who reported receiving feedback, treatment teachers indicated somewhat more positive perceptions about the information they received on their classroom practice but not about the information on their students' achievement.

Principals' Experiences

- Treatment principals reported receiving more than twice as many feedback sessions with ratings than control principals.
- Among those who reported receiving feedback, most principals in both treatment and control schools had positive perceptions about the feedback they received.

Findings About Teachers' Experiences

To measure teachers' experiences with performance feedback, we asked teachers in both treatment and control schools to complete a survey in the spring of the first year of study, usually at the beginning of the last of the four observation windows. In this section, we present survey-

[76] 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 are not included in the main study analyses, however, because by design they received limited feedback on classroom practice. They also received no feedback on student growth because student assessment data were not available in Kindergarten through grade 3.

based findings on the differences between treatment and control teachers in the amount and content of the performance feedback they received, and their perceptions of the feedback.

The Amount and Content of the Performance Feedback Teachers Received

The survey asked teachers to report on every instance in which they were observed and later received feedback during the first year of the study. The instances teachers reported on included observations for the purpose of evaluation as well as walkthroughs and informal observations (e.g., peer-to-peer observations). All but three treatment teachers (> 99 percent) and a large majority (88 percent) of control teachers reported receiving some form of feedback based on observations. In addition, a large majority (84 percent) of treatment teachers reported receiving feedback with ratings (see exhibit 5.1).

It was uncommon for control teachers to receive feedback with ratings. Less than half (39 percent) of the control teachers did so, which was consistent with the infrequent evaluations required by the study districts for nonprobationary teachers (see chapter 1).[77] The survey showed that 31.3 percent of nonprobationary teachers in control schools received feedback with ratings, compared with 68.8 percent of probationary teachers in control schools.[78]

[77] We identified probationary and nonprobationary teachers based on district policies that define the probationary period and teacher self-reported years of experience in the district.

[78] We also tested whether the treatment-control differences in teachers' experience with performance feedback differed by teachers' probationary status for all the other teacher survey measures presented in this chapter. The results were statistically significant for 4 of the 28 measures examined. (See a summary in exhibit I.9 in appendix I.) Where we did find a statistically significant difference in treatment effect by probationary status, the effect was larger for nonprobationary teachers than for probationary teachers.

Exhibit 5.1. Percentage of teachers who reported receiving ratings on their classroom practice, by treatment status

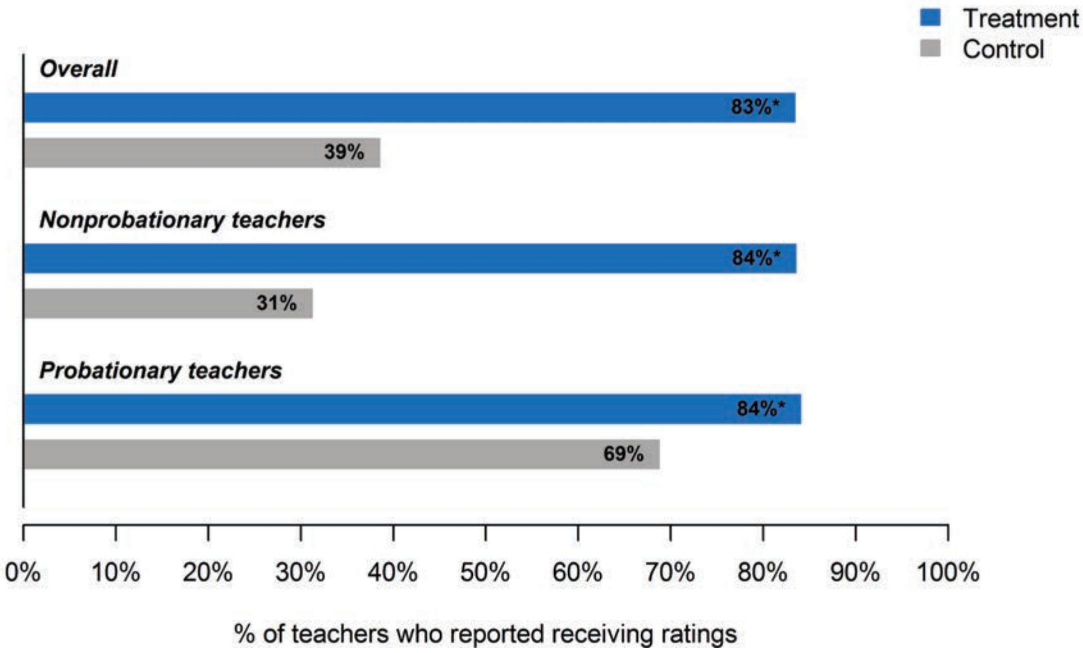


Exhibit Reads: Overall, 83 percent of treatment teachers and 39 percent of the control teachers reported receiving ratings on their classroom practice.

NOTE: Sample size = 127 schools (63 treatment and 64 control); 1,072 teachers (523 treatment and 549 control); 858 nonprobationary teachers (429 treatment and 429 control); and 213 probationary teachers (93 treatment and 120 control). The overall percentage of treatment teachers who reported receiving ratings is less than the percentage of probationary and nonprobationary teachers who reported receiving ratings because the probationary status for one of the treatment teachers is missing. The analyses were based on a two-level analysis (teachers within schools) controlling for random assignment blocks.

Statistically significant difference (*p* < .05, two-tailed) between the treatment and control groups is indicated by an asterisk (*) marking the treatment group mean. See exhibits I.1a, 1b, and 1c in appendix I for separate results for CLASS districts and FFT districts as well as results for K–3 teachers, respectively.

SOURCE: Spring 2013 Teacher Survey.

Treatment teachers reported receiving more than four times as many feedback sessions with ratings and a written narrative on their classroom practice as control teachers. The average treatment teacher reported receiving 3.0 instances of feedback sessions that included ratings and a written narrative, compared with 0.7 instances for the average control teacher (see exhibit 5.2). In addition, the total length of all oral feedback sessions received was 80 minutes for the average treatment teacher, compared to 18 minutes for the average control teacher.

Exhibit 5.2. Number of feedback instances and duration of feedback on classroom practice that an average teacher reported receiving, by treatment status

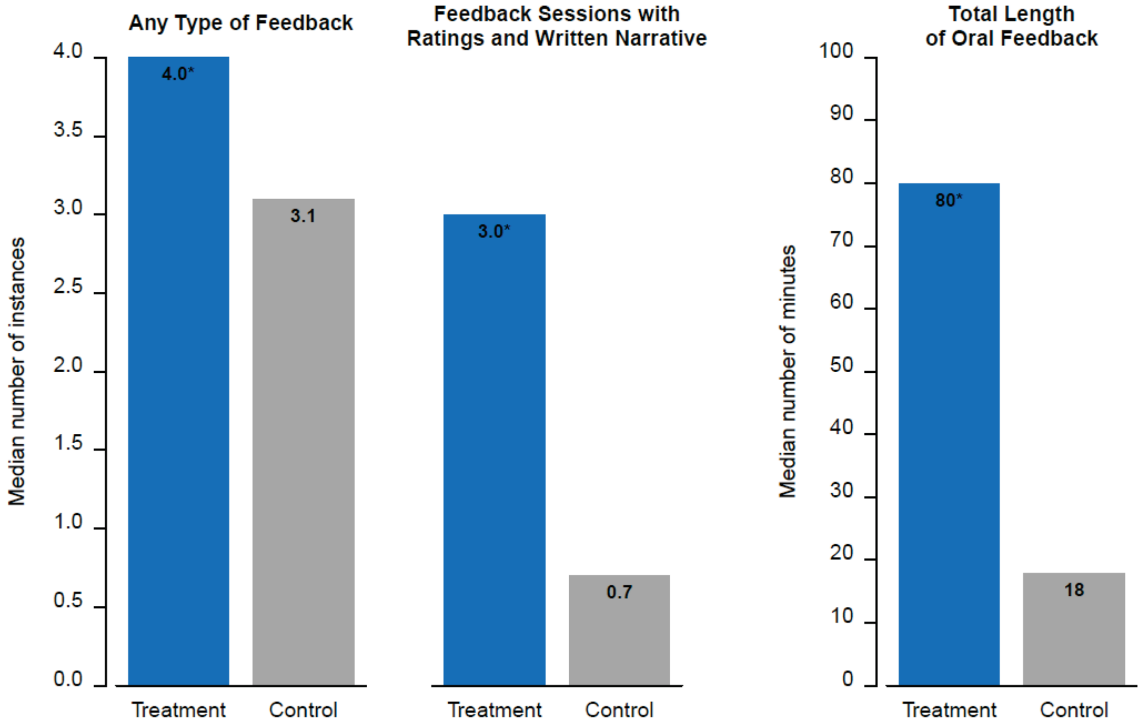


Exhibit Reads: The average treatment teacher reported receiving 4.0 instances of feedback on their classroom practice compared with 3.1 instances for control teachers.

NOTE: Sample size = 127 schools (63 treatment and 64 control) and 1,072 teachers (523 treatment and 549 control). The analyses were based on an aligned rank sum test with randomization inference about median difference between treatment and control groups (see appendix D for technical details). Statistically significant difference (p < .05, two-tailed) between the treatment and control groups is indicated by an asterisk (*) marking the treatment group median. See exhibits I.2a, 2b, and 2c in appendix I for separate results for CLASS districts and FFT districts as well as results for K–3 teachers, respectively.

SOURCE: Spring 2013 Teacher Survey.

Relative to control teachers, treatment teachers were more likely to report receiving feedback based on observations from observers not based at the teachers' schools. The intervention's measure of classroom practice was designed to provide teachers not only with observations by school administrators, but also observations by observers from outside their schools. Nearly all (94 percent) treatment teachers reported receiving feedback based on observations from their school administrators (typically the principals), compared with 86 percent of control teachers. In sharper contrast, treatment teachers were more than four times as likely to report receiving observation-based feedback from someone not from the teacher's school as control teachers (75 percent versus 16 percent),[79] which is likely to due to study-hired observers being used only in treatment schools.

[79] In the relevant question in the teacher survey, non-school-based observers excluded coaches or mentors.

More treatment than control teachers reported having discussions about CLASS/FFT-related areas of practice with someone who provided them with performance feedback. In theory, the intervention may shift the focus of feedback on teacher performance toward areas of classroom practice measured by CLASS and FFT. To test this theory, the teacher survey asked teachers whether they discussed specific areas with someone who provided feedback on their teaching, including areas related to CLASS/FFT as well as areas not related. Relative to control teachers, treatment teachers were more likely to report discussing four of the five areas of practice related to CLASS and FFT with someone who provided them with feedback (i.e., all but the dimension "behavior management"). Treatment teachers were no more likely to discuss areas not related to CLASS and FFT (see exhibit 5.3).

Exhibit 5.3. Percentage of teachers who reported discussing areas of classroom practice related to CLASS/FFT and areas not related, with someone who provided them with feedback during the school year, by treatment status

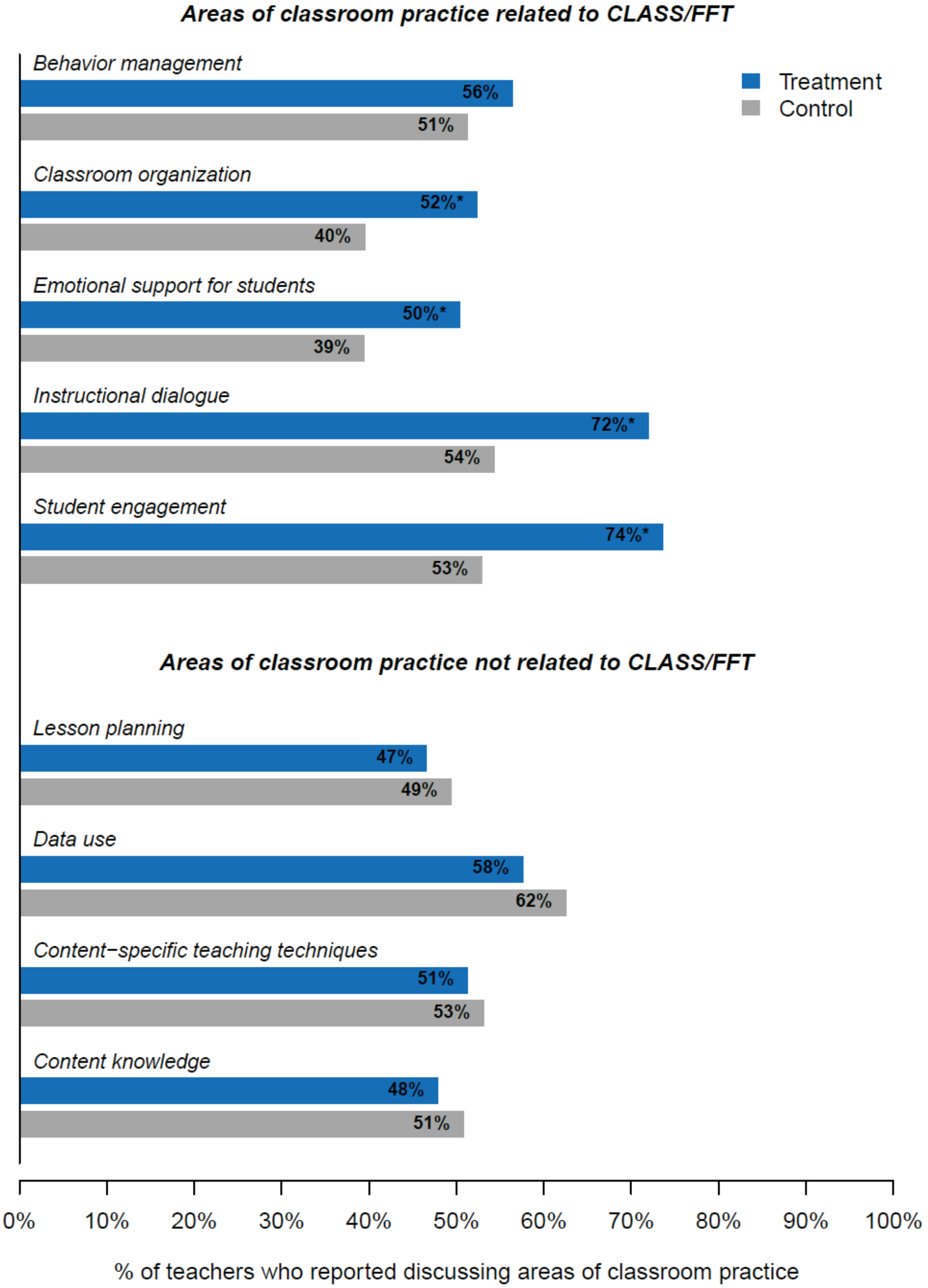


Exhibit Reads: Of treatment teachers, 56 percent reported discussing behavior management with someone who provided them with performance feedback during the school year, compared with 51 percent of control teachers.

NOTE: Sample size = 127 schools (63 treatment and 64 control) and 944–950 teachers (460–463 treatment and 484–488 control).

The analyses were based on a two-level analysis (teachers within schools) controlling for random assignment blocks. Statistically significant difference (*p* < .05, two-tailed) between the treatment and control groups is indicated by an asterisk (*) marking the treatment group mean. See exhibits I.3a, 3b, and 3c in appendix I for separate results for CLASS districts and FFT districts as well as results for K–3 teachers, respectively.

SOURCE: Spring 2013 Teacher Survey.

Relative to control teachers, treatment teachers were more likely to report receiving value-added scores and less likely to report receiving test scores for individual students or classroom average scores. The teacher survey asked teachers to report whether or not they received information on the achievement of their students, and it asked for separate responses about value-added scores, test scores of individual students, and class averages. Almost twice as many treatment teachers as control teachers reported receiving value-added scores based on the students that they taught during the previous school year (45 percent versus 24 percent; see exhibit 5.4). Fewer treatment teachers, however, reported receiving student achievement information for individual students they taught (64 percent versus 84 percent) or data on classroom averages (51 percent versus 62 percent). As a validity check, we compared treatment teachers' responses with electronic records indicating who had accessed their own value-added score in the online system, and found that 34 percent of the treatment teachers who reported receiving value-added scores did not access their student growth reports in the online system, and 17 percent of treatment teachers who reported not receiving value-added scores actually accessed their online student growth reports. Thus, many treatment teachers apparently did not understand that the intervention's measure of student growth provided value-added scores, raising questions about the validity of teachers' report of their receipt of value-added scores.

Exhibit 5.4. Percentage of teachers who reported receiving specific types of student achievement information, by treatment status

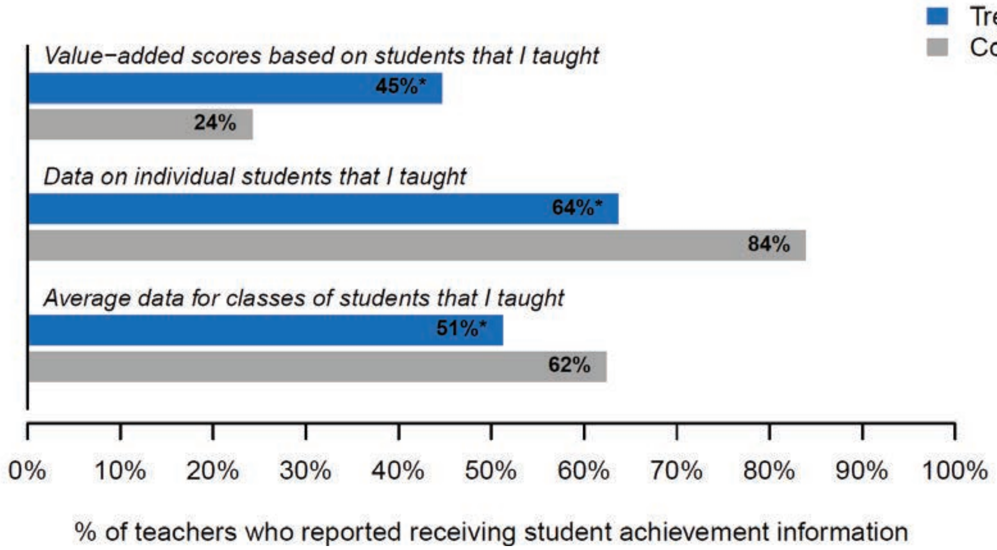


Exhibit Reads: Of treatment teachers, 45 percent reported receiving value-added scores based on the students they taught, compared with 24 percent of control teachers.

NOTE: Sample size = 127 schools (63 treatment and 64 control) and 1,073 teachers (519 treatment and 554 control). The analyses were based on a teacher-level regression controlling for random assignment blocks. Statistically significant difference (*p* < .05, two-tailed) between the treatment and control groups is indicated by an asterisk (*) marking the treatment group mean. See exhibits I.4a, 4b, and 4c in appendix I for separate results for CLASS districts and FFT districts as well as results for K–3 teachers, respectively. Findings about teachers' receipt of value-added scores should be interpreted with caution given that 34 percent of the treatment teachers who reported receiving value-added scores did not access their student growth reports in the study's online system, and 17 percent of treatment teachers who reported not receiving value-added scores actually accessed their online student growth reports.

SOURCE: Spring 2013 Teacher Survey.

Teachers' Perceptions About Performance Feedback and Rating Systems

Teachers who reported receiving feedback were asked about their perceptions, and these teachers are the basis of the analyses presented in this section. Because the teachers who reported receiving feedback are a selected subset of the full sample, these analyses should not be used to draw causal conclusions about the intervention's effects on teachers' perceptions; instead they are intended to describe and compare the perceptions of teachers in treatment and control schools who received feedback.[80]

In both treatment and control schools, a large majority of those who received feedback on classroom practice held positive views of the feedback, and on some measures, there was a statistically significant difference, with treatment teachers reporting more positive perceptions. We hypothesized that teachers' views about the feedback they received might influence any actions they might take in response. If teachers had negative views of the feedback, seeing it as unfair or vague, for example, they may have ignored it and continued their normal classroom practices. Therefore the survey asked teachers who had received feedback about the extent to which they agreed with five positive statements about the feedback they received during the year. Analyses focused on those who reported receiving feedback based on observations, which included all but three treatment teachers (over 99 percent) and 88 percent of control teachers in the full study sample.[81] Overall, treatment and control teachers were both likely to hold positive perceptions of their feedback. Even in the control group, more than three quarters (79 percent or more) reported positive perceptions. On three of the five measures, among those who received feedback on classroom practice, there was a statistically significant difference, with treatment teachers reporting more positive perceptions of the feedback they received. (See exhibit 5.5). For example, 87 percent of treatment teachers who received feedback indicated that their feedback included specific ideas for improvement, compared with 79 percent of control teachers, an 8 percentage point difference. That is the largest of the statistically significant differences.

[80] It would be possible to conduct the analyses based on the full sample, examining the percentage of teachers who *both* received feedback and reported positive perceptions. However, any difference between treatment and control teachers in this joint measure would largely reflect the group difference in the percentage of teachers who received feedback rather than the difference in teachers' perceptions about the feedback they received.

[81] To help assess the comparability of the treatment and control teachers who received feedback, we examined whether they differed in teaching experience and probationary status, two characteristics that might have been associated with whether teachers received feedback and also with their perceptions. No statistically significant differences were found (see exhibit I.10a in appendix I).

Exhibit 5.5. Percentage of teachers receiving feedback who agreed or strongly agreed with statements about that feedback, by treatment status

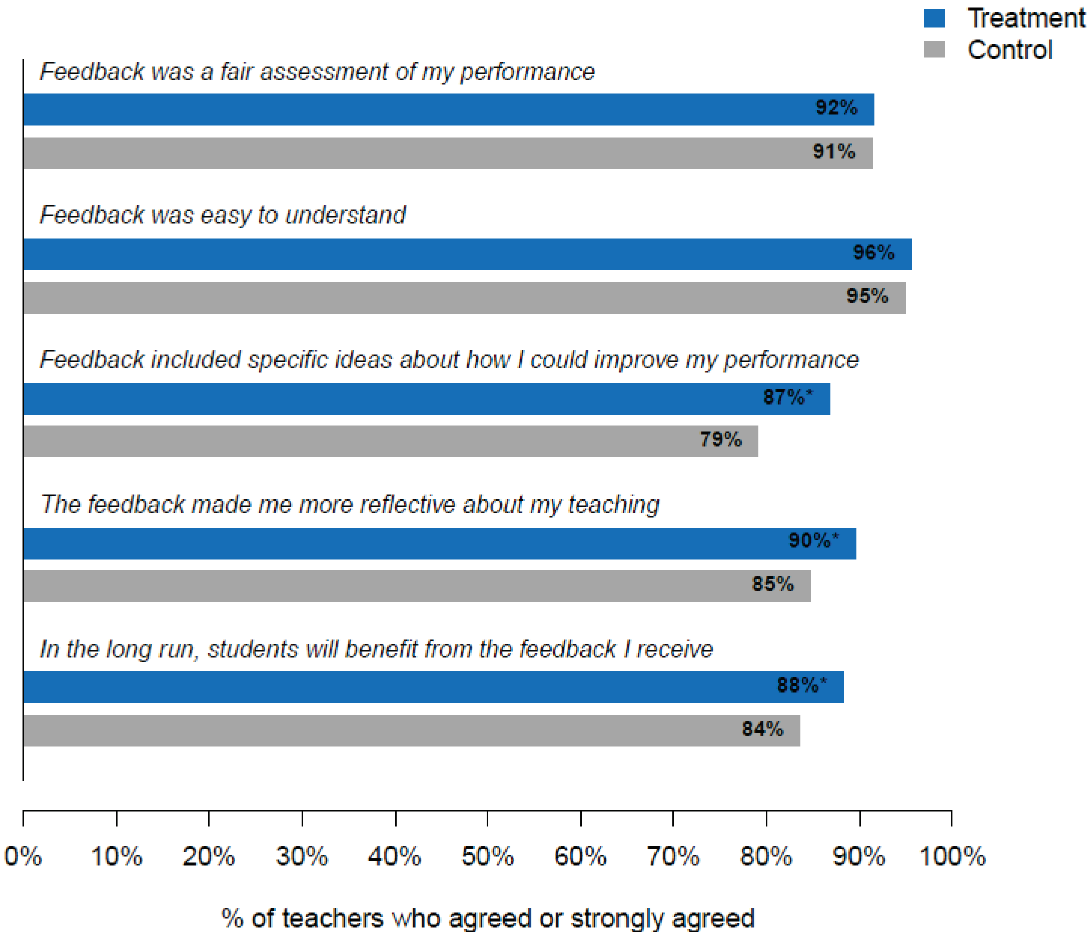


Exhibit Reads: Of treatment teachers who reported receiving feedback based on observations, 92 percent agreed or strongly agreed that the feedback they received was a fair assessment of their performance, compared with 91 percent of control teachers.

NOTE: Sample size = 127 schools (63 treatment and 64 control) and 1,004–1,008 teachers (519–512 treatment and 485–487 control). The analyses were based on a two-level analysis (teachers within schools) controlling for random assignment blocks. Statistically significant difference (*p* < .05, two-tailed) between the treatment and control groups is indicated by an asterisk (*) marking the treatment group mean. See exhibits I.5a, 5b, and 5c in appendix I for separate results for CLASS districts and FFT districts as well as results for K–3 teachers, respectively.

SOURCE: Spring 2013 Teacher Survey.

Most of the teachers who reported receiving observation-based ratings held positive views about the rating systems, but fewer treatment teachers than control teachers perceived the rating systems as fair to all teachers. Of all teachers in the full study sample, 83 percent of treatment teachers and 39 percent of control teachers reported receiving observation-based ratings. Overall, the majority in both groups who received such ratings believed that the rating systems did a good job distinguishing effective from ineffective teaching (78 percent and 81 percent, respectively) and provided accurate information about their teaching (77 percent and 82 percent, respectively) (see exhibit 5.6). In addition, a large majority (86 percent) of teachers in both treatment and control schools who received ratings reported that they had a clear idea of what the rating system viewed as good instruction. The majority of the teachers who reported receiving performance ratings also agreed or strongly agreed that the rating system was fair to all teachers, regardless of the characteristics of the teachers or their students. That view was held by fewer treatment than control teachers (67 percent versus 80 percent) who received an observation-based rating. However, the treatment-control difference may be influenced by the fact that the control teachers who received such ratings were a subset of less than half of the control teachers. They were also more likely than the control teachers who received no ratings to be novice and probationary teachers, because probationary teachers are typically observed more frequently than veteran, nonprobationary teachers, under districts' evaluation systems. Those attributes—being novice and on probationary status—may affect a teacher's perception of the fairness of the ratings given to them.[82]

[82] To help assess the comparability of the treatment and control teachers who received observation-based ratings, we examined whether they differed in teaching experience and probationary status, two characteristics that might have been associated with whether teachers received ratings, and also with their perceptions. Among teachers who reported receiving ratings, about half as many treatment teachers as control teachers had three or fewer years of teaching experience (12.8 percent vs. 26.3 percent) or were on probationary status (18.7 percent vs. 39.1 percent), differences that are statistically significant. (See exhibit I.10b in appendix I.) That result suggests caution in interpreting the comparison of treatment and control teachers who received ratings. However, we ran the analyses of perceptions for nonprobationary teachers only, and the results did not change except that there was no statistically significant difference for the item about fairness of the rating system. (See exhibit I.10d in appendix I.)

Exhibit 5.6. Percentage of teachers receiving observation-based ratings who agreed or strongly agreed with statements about the rating system used for the majority of the ratings they received, by treatment status

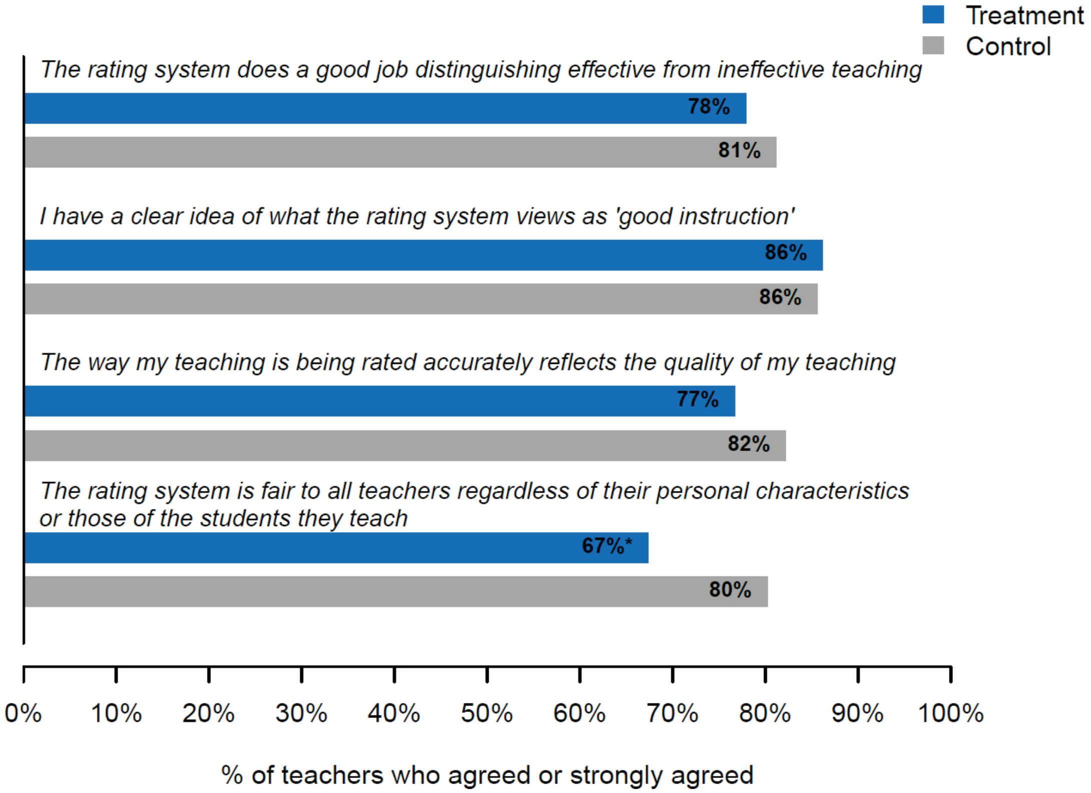


Exhibit Reads: Of treatment teachers who reported receiving observation-based ratings, 78 percent agreed or strongly agreed that the rating system did a good job distinguishing effective from ineffective teaching, compared with 81 percent of control teachers.

NOTE: Sample size = 122–123 schools (62 treatment and 60–61 control) and 631–639 teachers (419–428 treatment and 211–213 control). The analyses were based on a two-level analysis (teachers within schools) controlling for random assignment blocks. None of the differences between the treatment and the control groups were statistically significant at the .05 level (two-tailed). See exhibits I.6a, 6b, and 6c in appendix I for separate results for CLASS districts and FFT districts as well as results for K–3 teachers, respectively.

SOURCE: Spring 2013 Teacher Survey.

Treatment teachers were more likely than control teachers to perceive the student achievement information they received as difficult to understand, but they were more likely to perceive it as fair. Among treatment and control teachers who received information on the achievement of their students, most considered the information easy to understand, and yet most did not consider it fair as an assessment of teacher performance.[83] Compared with control teachers, fewer treatment teachers agreed or strongly agreed that the

[83] Analyses of teachers' perceptions of the student achievement information they received were restricted to the 85 percent of treatment teachers and 93 percent of control teachers in the full study sample who reported receiving student achievement information. To assess the comparability of these treatment and control teachers, we examined whether they differed in teaching experience and probationary status, two characteristics that might have been associated with teachers' perceptions. No statistically significant differences were found (see exhibit I.10c in appendix I). Findings from these analyses should be interpreted with caution given that some teachers might not have understood the survey item that asked about value-added scores. (See the discussion about teacher reports of receiving value-added scores preceding exhibit 5.4.)

student achievement information they received was easy to understand (78 percent versus 89 percent), which may reflect the complexity of value-added scores provided to treatment teachers as part of the intervention. At the same time, more treatment than control teachers perceived the information as fair (see exhibit 5.7).

Exhibit 5.7. Percentage of teachers receiving student achievement information who agreed or strongly agreed with statements about that information, by treatment status

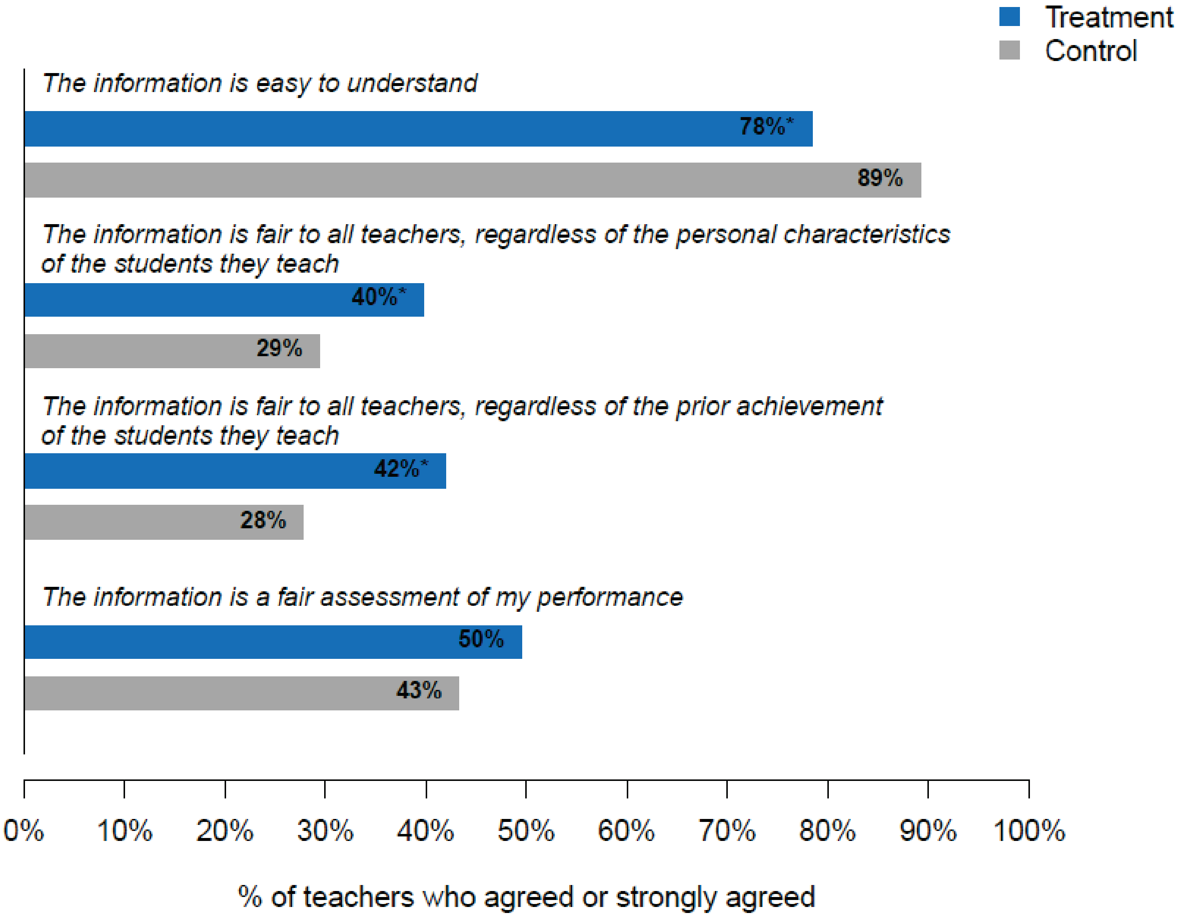


Exhibit Reads: Of those treatment teachers who reported receiving student achievement information, 40 percent agreed or strongly agreed that the student achievement information they received was fair to all teachers, regardless of the personal characteristics of the students they taught, compared with 29 percent of control teachers.

NOTE: Sample size = 127 schools (63 treatment and 64 control) and 949–953 teachers (437–439 treatment and 512 – 514 control). The analyses are based on a two-level linear regression model controlling for random assignment blocks. Statistically significant difference (*p* < .05, two-tailed) between the treatment and control groups is indicated by an asterisk (*) marking the treatment group mean. See exhibits I.7a, 7b, and 7c in appendix I for separate results for CLASS districts and FFT districts as well as results for K–3 teachers, respectively.

SOURCE: Spring 2013 Teacher Survey.

Findings About Principals' Experiences

To measure principals' experiences with performance feedback, we asked principals in both treatment and control schools to complete a survey in the spring of the first year of study, prior to the spring VAL-ED feedback session. In this section, we present survey-based findings on the differences between treatment and control principals in the amount and content of the performance feedback they received, and their perceptions of the feedback.

Amount and Content of the Performance Feedback Principals Received

Treatment principals reported receiving more feedback than control principals.
Compared with the average control principal, the average treatment principal reported receiving more instances of feedback (2.0 versus 1.4 instances) and more instances of oral feedback with ratings (1.0 versus 0.4 instances) (see exhibit 5.8). The average treatment principal also reported receiving a larger amount of oral feedback than did the average control principal (60 minutes versus 41 minutes).

Exhibit 5.8. Number of feedback instances and duration of oral feedback that principals reported receiving, by treatment status

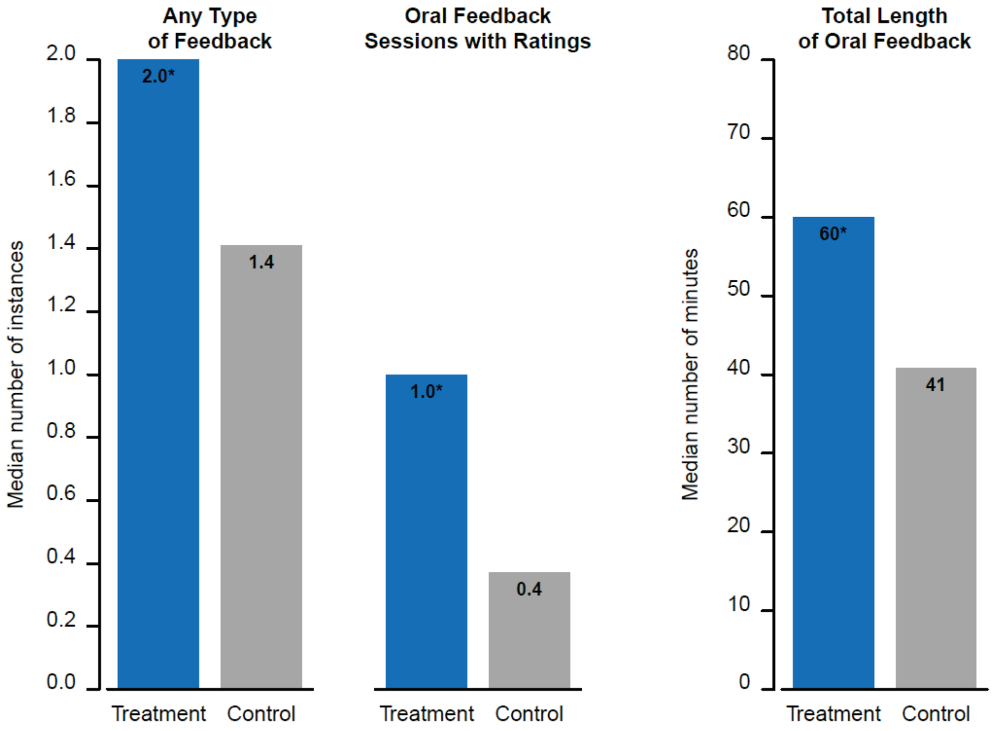


Exhibit Reads: The average treatment principal reported receiving 2.0 instances of feedback, compared with 1.4 instances for control principals.

NOTE: Sample size = 122 principals (61 treatment and 61 control). The analyses were based on an aligned rank sum test with randomization inference about median difference between treatment and control groups (see appendix D for technical details). Statistically significant difference (*p* < .05, two-tailed) between the treatment and control groups is indicated by an asterisk (*) marking the treatment group median.

SOURCE: Spring 2013 Principal Survey.

Treatment principals were no more likely than control principals to report discussing areas related to VAL-ED with their supervisors, except about parent/community issues. The principal survey asked the principals whether they discussed various areas with their supervisors, including areas aligned with the VAL-ED core components and areas unrelated to VAL-ED. Treatment principals were more likely than control principals to report discussing parent and community issues with their supervisors (70 percent versus 47 percent) but not other areas related to VAL-ED (see exhibit 5.9). Treatment and control principals were equally likely to report discussing areas unrelated to VAL-ED with their supervisors (e.g., making human resource decisions, managing nonpersonnel administrative issues, and student behavior and discipline) (see exhibit I.8 in appendix I).

Exhibit 5.9. Percentage of principals who reported discussing specific VAL-ED-related areas with their supervisors, by treatment status

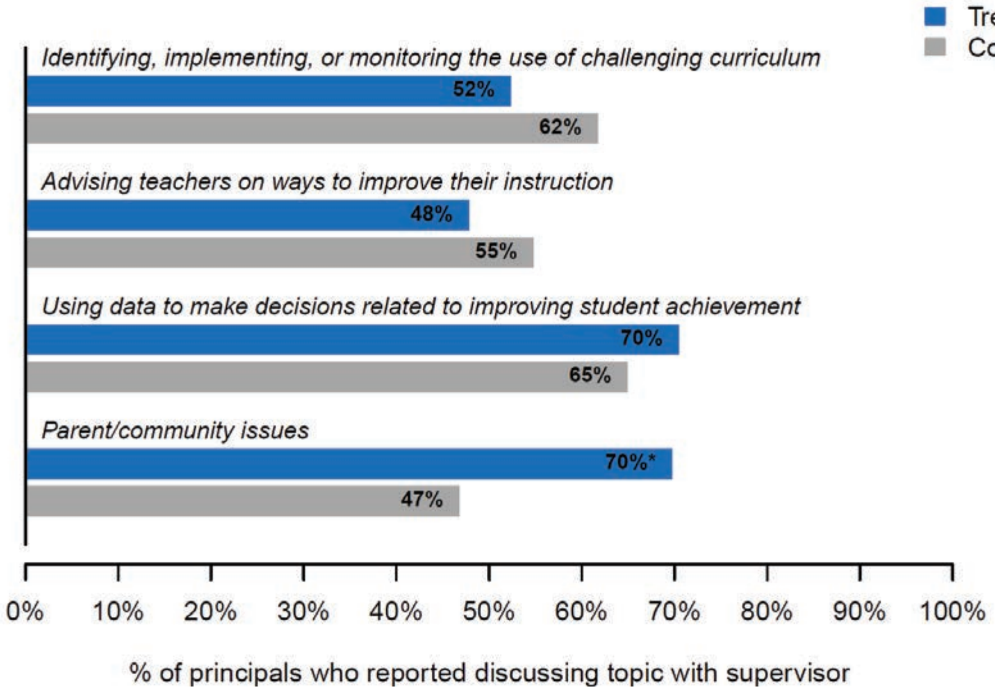


Exhibit Reads: Of treatment principals, 52 percent reported discussing with their supervisors in the area of identifying, implementing, or monitoring the use of challenging curriculum, compared with 62 percent of control principals.

NOTE: Sample size = 123 principals (61 treatment and 62 control). The analyses were based on a principal-level regression controlling for random assignment blocks. Statistically significant difference (*p* < .05, two-tailed) between the treatment and control groups is indicated by an asterisk (*) marking the treatment group mean.

SOURCE: Spring 2013 Principal Survey.

Principals' Perceptions About Performance Feedback

Principals who reported receiving feedback were asked about their perceptions, and these principals are the basis of the analyses presented in this section. As with the analyses of teachers' perceptions about feedback, the analyses of principal perceptions are based on a selected subset of the full sample, and should not be used to draw causal conclusions about the intervention's effects on principals' perceptions.

Among those who reported receiving feedback, most principals in both treatment and control schools had positive perceptions about the feedback they received. Eighty-seven percent of treatment principals and 58 percent of control principals reported receiving at least some feedback about their performance.[84] We surveyed this subset of principals about the fairness and specificity of the feedback. Over two thirds of these principals in both treatment and control schools (86 percent and 71 percent, respectively) agreed that the feedback they received was a fair assessment of their performance (see exhibit 5.10). The majority of these principals also agreed that the feedback they received contained specific ideas for improving their performance.

Exhibit 5.10. Percentage of principals receiving performance feedback who agreed or strongly agreed with statements about that feedback, by treatment status

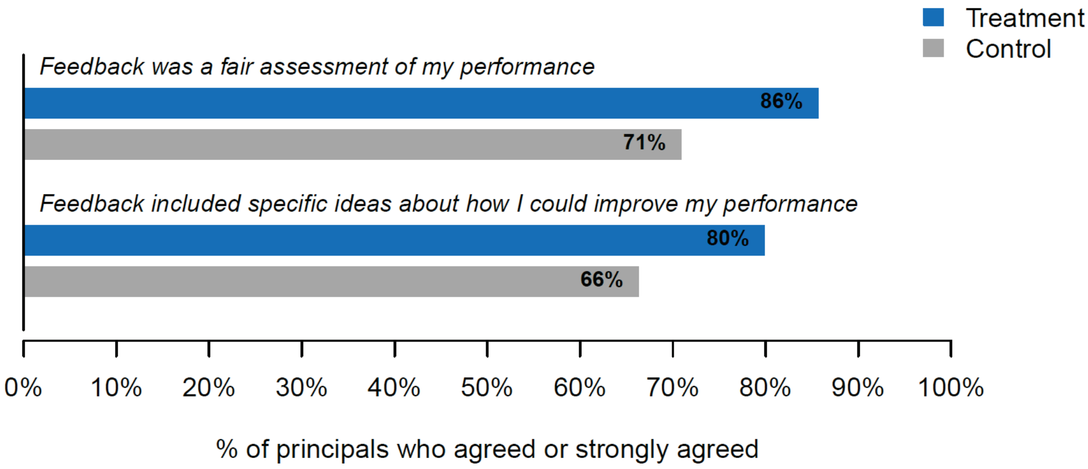


Exhibit Reads: Of treatment principals who reported receiving performance feedback from their supervisors, 86 percent agreed or strongly agreed that the feedback was a fair assessment of their performance, compared with 71 percent of control teachers.

NOTE: Sample size = 88 principals (53 treatment and 35 control). The analyses were based on a principal-level regression controlling for random assignment blocks. None of the differences between the treatment and the control groups were statistically significant at the .05 level (two-tailed).

SOURCE: Spring 2013 Principal Survey.

Summary

This chapter reported on the performance evaluation experiences of educators in the treatment and control schools. Treatment teachers received more feedback, including both classroom practice information and student growth information. The oral feedback based on classroom observations was of longer duration and more likely to include ratings and written narrative information. This suggests that treatment teachers received more in-depth and systematic feedback than control teachers, as intended. Among treatment and control teachers who received performance information, treatment teachers reported somewhat more positive perceptions about the information they received on their classroom practice, but not about the information on their

[84] To assess the comparability of these treatment and control principals, we examined whether they differed in their experience as a principal or as a teacher, which might have been associated with whether they received performance feedback, and also with their perceptions. No statistically significant differences were found (see exhibit I.10e in appendix I).

students’ achievement. Meanwhile, treatment principals reported receiving more instances of oral feedback with ratings of greater duration compared to control principals. However, treatment and control principals who received feedback were equally positive about the feedback they received. Further, they did not report differences in the topic areas in which they received feedback.

References

Albert, A., and Anderson, J.A. (1984). On the Existence of Maximum Likelihood Estimates in Logistic Regression Models. *Biometrika, 71*(1): 1–10.

Allen, J.P., Pianta, R.C., Gregory, A., Mikami, A.Y., and Lun, J. (2011). An Interaction-Based Approach to Enhancing Secondary School Instruction and Student Achievement. *Science, 333*: 1034–1037.

Allison, P. (2008). *Convergence Failures in Logistic Regression* (SAS Global Forum Paper 360-2008). Retrieved from http://www2.sas.com/proceedings/forum2008/360-2008.pdf.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Aspen Institute (2016). *Teacher Evaluation and Support Systems: A Roadmap for Improvement.* Washington, DC: Aspen Institute Education and Society Program.

Atwater, L.A. Brett, J.F., and Charles, A.C. (2007). Multisource Feedback: Lessons Learned and Implications for Practice. *Human Resources Management, 46*: 285–307.

Bill & Melinda Gates Foundation. (2012). *Gathering Feedback for Teaching: Combining High-Quality Observations With Student Surveys and Achievement Gains.* Seattle, WA: Author. Retrieved from http://www.metproject.org/downloads/MET_Gathering_Feedback_Research_Paper.pdf.

Bill & Melinda Gates Foundation. (2013). *Feedback for Better Teaching: Nine Principles for Using Measures of Effective Teaching*. Seattle, WA: Author. Retrieved from http://collegeready.gatesfoundation.org/wp-content/uploads/2015/05/MET_Feedback-for-Better-Teaching_Principles-Paper.pdf.

Blair, R.C., and Higgins, J.J. (1980). A Comparison of the Power of the Wilcoxon’s Rank-Sum Statistic to That of Student’s T Statistic Under Various Non-Normal Distributions. *Journal of Educational Statistics*, *5*(4): 309–335.

Casabianca, J.M., McCaffrey, D.F., Gitomer, D.H., Bell, C.A., Hamre, B.K., and Pianta, R.C. (2013). Effect of Observation Mode on Measures of Secondary Mathematics Teaching. *Educational and Psychological Measurement*, *73*(5): 757–783.

Chaplin, D., Gill, B., Thompkins, A., and Miller, H. (2014). *Professional Practice, Student Surveys, and Value-Added: Multiple Measures of Teacher Effectiveness in the Pittsburgh Public Schools* (REL 2014-024). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. Washington, DC: Regional Educational Laboratory Mid-Atlantic.

Chetty, R., Friedman, J.N., and Rockoff, J.E. (2014a). Measuring the Impacts of Teachers II: Evaluating Bias in Teacher Value-Added Estimates. *The American Economic Review*, *104*(9): 2593–2632.

Chetty, R., Friedman, J.N., and Rockoff, J.E. (2014b). Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood. *The American Economic Review*, *104*(9): 2633–2679.

Chiang, H., Wellington, A., Hallgren, K., Speroni, C., Herrmann, M., Glazerman, S., and Constantine, J. (2015). *Evaluation of the Teacher Incentive Fund: Implementation and Impacts of Pay-for-Performance After Two Years* (NCEE 2015-4020). U.S. Department of Education, Institute of Education Sciences. Washington, DC: National Center for Education Evaluation and Regional Assistance.

Collins, C., and Amrein-Beardsley, A. (2014). Putting Growth and Value-Added Models on the Map: A National Overview. *Teachers College Record, 116*: 1–34.

Condon, C., and Clifford, M. (2010). *Measuring Principal Performance: How Rigorous Are Commonly Used Principal Performance Assessment Instruments?* Naperville, IL: Learning Point Associates.

Chicago Consortium of School Research (2012). *2012 CPS My Voice, My School Teacher Survey Codebook*. Chicago, IL: Chicago Consortium of School Research. Retrieved from https://ccsr.uchicago.edu/sites/default/files/uploads/survey/2012%20CPS%20Teacher%20Survey%20Codebook.pdf.

Conway, J.M., and Huffcutt, A.I. (1997). Psychometric Properties of Multisource Performance Ratings: A Meta-Analysis of Subordinate, Supervisor, Peer, and Self-Ratings. *Human Performance*, *10*(4): 331–360.

Danielson, C. (2016). Charlotte Danielson on Rethinking Teacher Evaluation. *Education Week*. Retrieved from http://www.edweek.org/ew/articles/2016/04/20/charlotte-danielson-on-rethinking-teacher-evaluation.html.

Dee, T., and Wyckoff, J. (2013). *Incentives, Selection, and Teacher Performance: Evidence From IMPACT* (NBER Working Paper 19529). Cambridge, MA: National Bureau of Economic Research. Retrieved from http://www.nber.org.

Doran, H. (2014). Methods for Incorporating Measurement Error in Value-Added Models and Teacher Classifications. *Statistics and Public Policy, 1*(1): 114–119.

Fieller, E.C. (1954). Some Problems in Interval Estimation. *Journal of the Royal Statistical Society, Series B, 16*(2): 175–185.

Garet, M.S., Porter, A.C., Desimone, L.M., Birman, B.F., and Yoon, K.S. (2001). What Makes Professional Development Effective? Results From a National Sample of Teachers. *American Educational Research Journal, 38*(4): 915–945.

Glazerman, S., Isenberg, E., Dolfin, S., Bleeker, M., Johnson, A., Grider, M., and Jacobus, M. (2010). *Impacts of Comprehensive Teacher Induction: Final Results From a Randomized Controlled Study* (NCEE 2010-4028). U.S. Department of Education, Institute of Education Sciences. Washington, DC: National Center for Education Evaluation and Regional Assistance.

Glazerman, S., and Saifullah. A. (2010). *An Evaluation of the Teacher Advancement Program (TAP) in Chicago: Year Two Impact Report*. Mathematica Policy Research.

Goe, L., Bell, C., and Little, O. (2008). *Approaches to Evaluating Teacher Effectiveness.* Washington, DC: National Comprehensive Center for Teacher Quality.

Goldhaber, D., and Hansen, M. (2013). Is It Just a Bad Class? Assessing the Long-Term Stability of Estimated Teacher Performance. *Economica*, *80*: 589–612. doi: 10.1111/ecca.12002.

Goldring, E., Carvens, X., Murphy, J., Porter, A., Elliott, S., and Carson, B. (2009). The Evaluation of Principals: What and How Do States and Urban Districts Assess Leadership? *Elementary School Journal, 110*(1): 19–39.

Grossman, P., Loeb, S., Cohen, J., and Wyckoff, J. (2013). Measure for Measure: The Relationship Between Measures of Instructional Practice in Middle School English Language Arts and Teachers' Value-Added Scores. *American Journal of Education, 119*(3): 445–470.

Hanushek, E.A., and Rivkin, S.G. (2010). Generalizations About Using Value-Added Measures of Teacher Quality. *The American Economic Review*, 267–271.

Hill, H.C., Kapitula, L., and Umland, K. (2011). A Validity Argument Approach to Evaluating Teacher Value-Added Scores. *American Educational Research Journal*, *48*(3): 794–831.

Ho, A.D., and Kane, T.J. (2013). *The Reliability of Classroom Observations by School Personnel.* Seattle, WA: Bill & Melinda Gates Foundation.

Hodges, J.L., and Lehmann, E. (1962). Rank Methods for Combination of Independent Experiments in Analysis of Variance. *The Annals of Mathematical Statistics, 33*(2): 482–497.

Kane, T.J., Taylor, E.S., Tyler, J.H., and Wooten, A.L. (2011). Identifying Effective Classroom Practices Using Student Achievement Data. *Journal of Human Resources*, *46*(3): 587–613.

Kane, T.J., McCaffrey, D.F., Miller, T., and Staiger, D.O. (2013). *Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment.* Seattle, WA: Bill & Melinda Gates Foundation.

Kane, T.J., and Staiger, D.O. (2008). *Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation* (No. w14607). Cambridge, MA: National Bureau of Economic Research.

Kane, T.J., and Staiger, D.O. (2012). *Gathering Feedback for Teaching: Combining High-Quality Observations With Student Surveys and Achievement Gains.* Seattle, WA: Bill & Melinda Gates Foundation.

Kitchen, C.M. (2009). Nonparametric vs. Parametric Tests of Location in Biomedical Research. *American Journal of Ophthalmology, 147*(4): 571–572.

Lipscomb, S., Terziev, J., and Chaplin, D. (2015). *Measuring Teachers' Effectiveness: A Report from Phase 3 of Pennsylvania's Pilot of the Framework for Teaching.* Princeton, NJ: Mathematica Policy Research.

Mashburn, A.J., Downer, J.T., Hamre, B.K., Justice, L.M., and Pianta, R.C. (2010). Consultation for Teachers and Children's Language and Literacy Development During Pre-Kindergarten. *Applied Developmental Science, 14*(4): 179–196.

McCaffrey, D.F., Sass, T.R., Lockwood, J.R., and Mihaly, K. (2009). *The Inter-Temporal Variability of Teacher Effect Estimates* (Working Paper 2009-03). Vanderbilt University. Nashville, TN: National Center on Performance Incentives.

Mihaly, K., McCaffrey, D.F., Staiger, D.O., and Lockwood, J.R. (2013). *A Composite Estimator of Effective Teaching*. Santa Monica, CA: RAND Corporation.

Porter, A.C., Goldring, E., Elliott, S.N., Murphy, J., Polikoff, M.S., and Cravens, X.C. (2008). *Setting Performance Standards VAL-ED Assessment of Principal Leadership* (ERIC Document No. ED505799). Retrieved from http://files.eric.ed.gov/fulltext/ED505799.pdf.

Porter, A.C., Polikoff, M.S., Goldring, E., Murphy, J., Elliott, S.N., and May, H. (2010). Investigating the Validity and Reliability of the Vanderbilt Assessment of Leadership in Education. *Elementary School Journal, 111*(2): 282–313.

Puma, M.J., Olsen, R.B., Bell, S.H., and Price, C. (2009). *What to Do When Data Are Missing in Group Randomized Controlled Trials* (NCEE 2009–0049). U.S. Department of Education, Institute of Education Sciences. Washington, DC: National Center for Education Evaluation and Regional Assistance. Retrieved from http://files.eric.ed.gov/fulltext/ED511781.pdf.

Raudenbush, S., and Jean, M. (2012). *How Should Educators Interpret Value-Added Scores?* Carnegie Knowledge Network. Stanford, CA: Carnegie Foundation for the Advancement of Teaching. Retrieved from http://www.carnegieknowledgenetwork.org/wp-content/uploads/2012/10/CKN_2012-10_Raudenbush.pdf.

Rockoff, J.E., Staiger, D.O., Kane, T.J., and Taylor, E.S. (2011). *Information and Employee Evaluation: Evidence From a Randomized Intervention in Public Schools* (NBER Working Paper No. 16240). Cambridge, MA: National Bureau of Economic Research.

Sawchuk, S. (2016, January 6). Law Could Spur Changes in Teacher Requirements. *Education Week.*

Schochet, P.Z. and Chiang, H.S. (2013). What Are Error Rates for Classifying Teacher and School Performance Using Value-Added Models? *Journal of Educational and Behavioral Statistics, 38*(2): 142–171.

Sebastian, J., and Allensworth, E. (2012). The Influence of Principal Leadership on Classroom Instruction and Student Learning: A Study of Mediated Pathways to Learning. *Educational Administration Quarterly, 48*(4): 626–663.

Shavelson, R.J., and Webb, N.M. (1991). *Generalizability Theory: A Primer* (Vol. 1). Newbury Park, CA: Sage.

Smither, J.W., London, M., and Reilly, R.R. (2005). Does Performance Improve Following Multisource Feedback? A Theoretical Model, Meta-Analysis, and Review of Empirical Findings. *Personnel Psychology, 58*: 33–66.

Stecher, B., Garet, M.S., Hamilton, L.S., Steiner, E.D., Robyn, A., Porier, J., Holzman, D., Fulbeck, E.S., Chambers, J., and de los Reyes, I.B. (2016). *Improving Teaching Effectiveness: Implementation. The Intensive Partnerships for Effective Teaching Through 2013–2014*. Santa Monica, CA: RAND.

Steinberg, M., and Sartain, L. (in press). Does Teacher Evaluation Improve School Performance? Experimental Evidence from Chicago's Excellence in Teaching Project. *Education Finance and Policy*.

Taylor, E.S., and Tyler, J.H. (2012). The Effect of Evaluation on Teacher Performance. *American Economic Review, 102*(7): 3628–3651.

U.S. Department of Education. (2012). *ESEA Flexibility*. Washington, DC: Author. Retrieved from http://www2.ed.gov/policy/eseaflex/approved-requests/flexrequest.doc.

U.S. Department of Labor, Employment and Training Administration. (2006). *Testing and Assessment: A Guide to Good Practices for Workforce Investment Professionals*. Washington, DC: Author.

Viswesvaran, C., Ones, D.S., and Schmidt, F.L. (1996). Comparative Analysis of the Reliability of Job Performance Ratings, *Journal of Applied Psychology*, *81*(5): 557–572.

Webb, N.M., Shavelson, R.J., and Haertel, E.H. (2006). Reliability Coefficients and Generalizability Theory. *Handbook of Statistics*, *26*(4): 81–124.

Weisberg, D., Sexton, S., Mulhern, J., and Keeling, D. (2009). *The Widget Effect: Our National Failure to Acknowledge and Act on Differences in Teacher Effectiveness.* Brooklyn, NY: The New Teacher Project.

Weiner, R. (2016). Three Strategies to Improve Teacher Evaluation. *Education Week*. Retrieved from http://www.edweek.org/ew/articles/2016/04/27/three-strategies-to-improve-teacher-evaluation.html.

Whitehurst, G.J., Chingos, M.M., and Lindquist, K.M. (2014). *Evaluating Teachers With Classroom Observations: Lessons Learned in Four Districts*. The Brookings Institution. Washington, DC: Brown Center on Education Policy.

Appendixes

Appendix A. Details About the Study Sample

This appendix presents additional details about the study sample. The first section compares the characteristics of the study sample with the characteristics of broader populations (i.e., public schools in similarly sized districts and the national population of public schools). The second section presents baseline equivalence information for CLASS districts and FFT districts separately.

Similarity of the Study Sample to Broader Populations

To provide a broader frame of reference for the characteristics of the study sample, we compared the background characteristics of study schools with the characteristics of schools in similarly sized districts (i.e., districts with at least 20 elementary and middle schools) and schools in the national population. The results for elementary schools are presented in exhibit A.1; the results for middle schools are presented in exhibit A.2.

Exhibit A.1. Background characteristics for elementary schools in the study sample, elementary schools in similarly sized districts, and the national population, 2011–12

	Elementary schools in		
School characteristic	**Study sample**	**Similarly sized districts**	**National population**
Geographic region (percentage of schools)			
Northeast	0.0	8.8*	16.7*
South	41.7	45.8	33.0
Midwest	27.1	12.8*	24.9
West	31.3	27.6	23.1
Urbanicity (percentage of schools)			
Urban	60.4	52.4	25.7*
Suburban	17.7	33.1*	30.8*
Rural	21.9	14.6	43.3*
Title I status (percentage of schools)	75.0	73.9	78.8
Free or reduced-price lunch (school average percentage of students)	39.6	60.8*	52.9*
Minority/non-White (school average percentage of students)	57.4	66.3*	45.6*
Female (school average percentage of students)	48.4	48.3	48.3
Total school enrollment	479.2	545.3*	456.1
Number of full-time equivalent teachers (all grades)	29.0	32.6*	27.9
Number of schools	**96**	**18,481**	**49,507**

NOTE: “Similarly sized districts” are districts with at least 20 elementary and middle schools. Percentage values for characteristics with multiple categories may not sum to 100 because of rounding. Differences between study schools and schools in similarly sized districts or the national population were tested using *t* tests. Two-tailed statistical significance at the *p* < .05 level is indicated by an asterisk (*).

SOURCE: 2011–12 Common Core of Data.

Exhibit A.2. Background characteristics for middle schools in the study sample, middle schools in similarly sized districts, and the national population, 2011–12

	Middle schools in		
School characteristic	**Study sample**	**Similarly sized districts**	**National population**
Geographic region (percentage of schools)			
Northeast	0.0	8.5*	16.4*
South	45.2	51.9	35.5
Midwest	25.8	9.7	26.2
West	29.0	24.7	20.1
Urbanicity (percentage of schools)			
Urban	64.5	47.1	19.2*
Suburban	12.9	33.9*	29.7*
Rural	22.6	19.0	51.0*
Title I status (percentage of schools)	58.1	67.4	72.8
Free or reduced-price lunch (school average percentage of students)	41.6	56.5*	48.6
Minority/non-White (school average percentage of students)	57.2	63.0	40.6*
Female (school average percentage of students)	48.2	48.5	48.6
Total school enrollment	651.0	775.0*	582.7
Number of full-time equivalent teachers (all grades)	43.8	45.9	36.4*
Number of schools	**31**	**4,563**	**15,514**

NOTE: “Similarly sized districts” are districts with at least 20 elementary and middle schools. Percentage values for characteristics with multiple categories may not sum to 100 because of rounding. Differences between study schools and schools in similarly sized districts or the national population were tested using *t* tests. Two-tailed statistical significance at the *p* < .05 level is indicated by an asterisk (*).

SOURCE: 2011–12 Common Core of Data.

Exhibit A.3. Background characteristics for schools in CLASS and FFT districts, 2011–12

School characteristic	CLASS districts	FFT districts
Geographic region (percentage of schools)		
Northeast	0.0	0.0
South	63.5	21.9
Midwest	36.5	17.2
West	0.0	60.9
Urbanicity (percentage of schools)		
Urban	60.3	62.5
Suburban	30.2	†
Rural	9.5	†
Title I status (percentage of schools)	81.0	60.9
Free or reduced-price lunch (school average percentage of students)	36.2	43.9
Minority/non-White (school average percentage of students)	72.1	42.9
Female (school average percentage of students)	48.5	48.3
Total school enrollment	632.0	411.9
Number of full-time equivalent teachers (all grades)	38.9	26.3
Number of schools	**63**	**64**

NOTE: Percentage values for characteristics with multiple categories may not sum to 100 because of rounding.

† Figures suppressed due to small number of FFT districts in suburban areas.

SOURCE: 2011–12 Common Core of Data.

Supplemental Baseline Equivalence Test Results

This section presents the results of baseline equivalence tests that compare the background characteristics of schools, principals, teachers, and students between the treatment group and the control group for CLASS districts and FFT districts separately. The results for CLASS districts are provided in exhibits A.4a, A.5a, A.6a, and A.7a; the results for FFT districts are provided in exhibits A.4b, A.5b, A.6b, and A.7b.

Exhibit A.4a. School background characteristics in CLASS districts, by study group

Characteristic	Treatment group	Control group	Estimated difference	*p* value
Title I status (percentage)	80.6	82.1	-1.4	.641
Total school enrollment	623.5	627.0	-3.4	.787
Number of full-time equivalent teachers	39.3	38.8	0.4	.587
Percentage eligible for free and reduced-price lunch	36.7	36.2	0.5	.484
Percentage minority	73.5	72.8	0.7	.277
Percentage female	49.1	48.5	0.6*	.013
Number of schools	**31**	**32**		

NOTE: The analyses are based on an ordinary least-squares (OLS) regression model controlling for random assignment blocks. The treatment group means are unadjusted means; the control group means were computed by subtracting the estimated group differences from the unadjusted treatment group means. *p* Values are based on *t* tests. Two-tailed statistical significance at the *p* < .05 level is indicated by an asterisk (*).

SOURCE: 2011–12 Common Core of Data.

Exhibit A.4b. School background characteristics in FFT districts, by study group

Characteristic	Treatment group	Control group	Estimated difference	*p* value
Title I status (percentage)	59.4	61.4	-2.0	.549
Total school enrollment	402.0	401.3	0.7	.944
Number of full-time equivalent teachers	25.2	25.4	-0.2	.750
Percentage eligible for free and reduced-price lunch	43.2	44.6	-1.3	.263
Percentage minority	41.7	43.4	-1.7	.190
Percentage female	47.8	48.3	-0.5	.107
Number of schools	**32**	**32**		

NOTE: The analyses are based on an OLS regression model controlling for random assignment blocks. The treatment group means are unadjusted means; the control group means were computed by subtracting the estimated group differences from the unadjusted treatment group means. *p* Values are based on *t* tests. Two-tailed statistical significance at the *p* < .05 level is indicated by an asterisk (*).

SOURCE: 2011–12 Common Core of Data.

Exhibit A.5a. Principal background characteristics in CLASS districts, fall 2012, by study group

Characteristic	Treatment group	Control group	Estimated difference	*p* value
Years of experience in district				
Mean number of years	16.4	20.6	-4.2	.093
Three years or fewer (percentage)	†	†	11.0	.056
Four to 10 years (percentage)	†	†	-12.0	.159
Eleven to 20 years (percentage)	41.9	33.2	8.8	.498
More than 20 years (percentage)	35.5	43.2	-7.7	.568
Master's degree or higher (percentage)	†	†	-4.3	.486
Number of principals	**31**	**32**		

NOTE: The analyses are based on an OLS regression model controlling for random assignment blocks. The treatment group means are unadjusted means; the control group means were computed by subtracting the estimated group differences from the unadjusted treatment group means. *p* Values are based on *t* tests. Two-tailed statistical significance at the *p* < .05 level is indicated by an asterisk (*).

† Figures suppressed due to small number of control principals with three or fewer years of experience and small number of principals without a Master's degree or higher.

SOURCE: Fall 2012 District Archival Records.

Exhibit A.5b. Principal background characteristics in FFT districts, fall 2012, by study group

Characteristic	Treatment group	Control group	Estimated difference	*p* value
Years of experience in district				
Mean number of years	11.8	12.2	-0.3	.854
Three years or fewer (percentage)	25.0	15.1	9.9	.327
Four to 10 years (percentage)	25.0	44.3	-19.3	.076
Eleven to 20 years (percentage)	25.0	18.4	6.6	.506
More than 20 years (percentage)	25.0	22.1	2.9	.733
Master's degree or higher (percentage)	100.0	100.0	0.0	1.000
Number of principals	**32**	**32**		

NOTE: The analyses are based on an OLS regression model controlling for random assignment blocks. The treatment group means are unadjusted means; the control group means were computed by subtracting the estimated group differences from the unadjusted treatment group means. *p* Values are based on *t* tests. Two-tailed statistical significance at the *p* < .05 level is indicated by an asterisk (*).

SOURCE: Fall 2012 District Archival Records.

Exhibit A.6a. Teacher background characteristics in CLASS districts, fall 2012, by study group

Characteristic	Treatment group	Control group	Estimated difference	*p* value
Years of experience in district				
Mean number of years	10.5	9.8	0.6	.255
Three years or fewer (percentage)	20.1	23.4	-3.4	.186
Four to 10 years (percentage)	41.3	38.1	3.1	.224
Eleven to 20 years (percentage)	23.3	24.6	-1.4	.545
More than 20 years (percentage)	15.4	14.1	1.3	.511
Master's degree or higher (percentage)	31.7	33.8	-2.1	.283
Number of teachers	**718**	**745**		

NOTE: The analyses are based on a two-level linear regression model controlling for random assignment blocks. The treatment group means are unadjusted means, and the control group means were computed by subtracting the estimated group differences from the unadjusted treatment group means. *p* Values are based on *t* tests. Two-tailed statistical significance at the *p* < .05 level is indicated by an asterisk (*).

SOURCE: Fall 2012 District Archival Records.

Exhibit A.6b. Teacher background characteristics in FFT districts, fall 2012, by study group

Characteristic	Treatment group	Control group	Estimated difference	*p* value
Years of experience in district				
Mean number of years	8.7	10.8	-2.1*	.008
Three years or fewer (percentage)	31.9	25.5	6.4	.090
Four to 10 years (percentage)	35.1	33.2	1.9	.638
Eleven to 20 years (percentage)	23.3	24.1	-0.8	.823
More than 20 years (percentage)	9.6	17.2	-7.5*	.008
Master's degree or higher (percentage)	55.7	54.7	1.0	.792
Number of teachers	**509**	**509**		

NOTE: The analyses are based on a two-level linear regression model controlling for random assignment blocks. The treatment group means are unadjusted means, and the control group means were computed by subtracting the estimated group differences from the unadjusted treatment group means. *p* Values are based on *t* tests. Two-tailed statistical significance at the *p* < .05 level is indicated by an asterisk (*).

SOURCE: Fall 2012 District Archival Records.

Exhibit A.7a. Student background characteristics in CLASS districts, fall 2012, by study group (Grades 4–8)

Characteristic	Treatment group	Control group	Estimated difference	*p* value
Students eligible for free or reduced-price lunch (percentage)	66.0	66.0	0.0	.979
Race/ethnicity (percentage)				
White	28.0	29.0	-1.0	.233
Black or African American	3.5	3.9	-0.5	.240
Hispanic	64.6	63.4	1.2	.073
Asian/Pacific Islander	3.8	3.6	0.1	.705
Other	0.2	0.2	0.0	.971
Female (percentage)	50.0	48.0	2.0*	.008
English language learners (percentage)	27.1	28.9	-1.8	.443
Students with disabilities (percentage)	9.8	6.0	3.8*	.027
Student achievement on state assessment (standardized)				
2011–12 Mathematics achievement	0.006	-0.010	0.016	.776
2011–12 Reading/ELA achievement	-0.012	0.042	-0.054	.170
Number of students	**9,305**	**10,086**		

NOTE: The analyses are based on a three-level linear regression model controlling for random assignment blocks. The treatment group means are unadjusted means, and the control group means were computed by subtracting the estimated group differences from the unadjusted treatment group means. *p* Values are based on *t* tests. Two-tailed statistical significance at the *p* < .05 level is indicated by an asterisk (*).

SOURCE: Fall 2012 District Archival Records.

Exhibit A.7b. Student background characteristics in FFT districts, fall 2012, by study group (Grades 4–8)

Characteristic	Treatment group	Control group	Estimated difference	*p* value
Students eligible for free or reduced-price lunch (percentage)	54.6	57.3	-2.8	.237
Race/Ethnicity (percentage)				
White	59.9	56.8	3.1	.168
Black or African American	2.7	2.9	-0.2	.813
Hispanic	31.5	33.6	-2.1	.389
Asian/Pacific Islander	1.3	1.4	-0.1	.660
Other	4.7	5.5	-0.8	.653
Female (percentage)	48.2	48.3	-0.1	.944
English language learners (percentage)	4.5	5.5	-0.9	.520
Students with disabilities (percentage)	13.5	13.6	-0.1	.936
Student achievement on state assessment (standardized)				
2011–12 Mathematics achievement	-0.024	0.000	-0.024	.643
2011–12 Reading/ELA achievement	-0.045	0.010	-0.056	.263
Number of students	**6,246**	**7,222**		

NOTE: The analyses are based on a three-level linear regression model controlling for random assignment blocks. The treatment group means are unadjusted means, and the control group means were computed by subtracting the estimated group differences from the unadjusted treatment group means. *p* Values are based on *t* tests. Two-tailed statistical significance at the *p* < .05 level is indicated by an asterisk (*).

SOURCE: Fall 2012 District Archival Records.

Study-Hired Observer Characteristics

Exhibit A.8. Study-hired observer background characteristics, fall 2012, by classroom observation provider

Characteristic	All Districts	CLASS	FFT
Years of experience			
Mean number of years teaching	18.2	21.4	15.4
Mean number of years teaching English/Language Arts	6.5	8.4	4.8
Mean number of years teaching mathematics	9.1	10.4	7.9
Mean number of years as an instructional coach	3.4	3.6	3.3
Mean number of years as a school administrator	5.6	6.4	4.9
Mean number of years as a district administrator	1.5	1.2	1.8
Master's degree or higher (percentage)	94.0	†	†
Main activity in year prior to study (2011-12)			
Teaching (percentage)	28.4	32.3	25.0
School administrator (percentage)	11.9	†	†
Other position (percentage)	43.3	†	†
Retired or unemployed (percentage)	16.4	22.6	11.1
Had formal training on observing teachers prior to the study (percentage)	71.6	71.0	72.2
Had formal training on providing feedback to teachers on their instructional practice prior to the study (percentage)	70.0	66.7	72.2
Number of study-hired observers	**67**	**31**	**36**

NOTE: Percentage values for characteristics with multiple categories may not sum to 100 because of rounding. The sample size for the years of experience characteristics ranges from 64-67 (30-31 in CLASS districts, 34-36 in FFT districts).

† Figures suppressed due to small number of observers without a Master's degree or higher and small number of observers who were school administrators in the year prior to the study.

SOURCE: 2012 Observer Information Sheet.

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

Appendix B. Details About Data Collection

This appendix provides details on the data collection activities that occurred during the first year of the study. Information is provided on four types of data collected: intervention implementation, educator performance, educators' experiences with performance feedback and initial outcomes, and the characteristics of study participants.

Intervention Implementation

To examine the extent to which the intervention was implemented as intended, we collected data from a variety of sources at different times throughout the year, as shown in exhibit B.1 and described in more detail next.

Exhibit B.1. Collection of intervention implementation data in the first year of the study (2012–13)

Data	2012–13			
	Jul.–Sep.	**Oct.–Dec.**	**Jan.–Mar.**	**Apr.–Jun.**
Event delivery and participation measures	Summer			
Observer information sheets and certification results	Summer			
Study-hired observer questionnaire				End of year
CLASS/FFT online system records	Throughout school year			
VAL-ED online system records		November		April
AIR online system records				End of year
District interviews				**End of year**

Event delivery and participation measures. We collected data on the fidelity of the delivery of and participation in key intervention events through in-person visits. A member of the implementation team attended each orientation and training event to collect attendance sheets and the agenda/schedule, and recorded the actual length of each section on the agenda. For webinars, the implementation team member collected the same information through the Web.

Observer information sheets and certification results. The implementation team reserved at least 10 minutes during the observer training for observers (principals and study-hired observers) to complete a short information sheet to gather information such as their degree(s); years of experience as a teacher, administrator, and/or evaluator; and prior observation experience. Shortly after the training, we collected observer certification test results for each observer using the provider online system.

Study-hired observer questionnaire. At the end of the first year of the study, a questionnaire was administered to each study-hired observer, focusing on time spent performing their duties, their practices in conducting feedback sessions, their self-confidence as raters and givers of feedback, and their general beliefs about scoring observations and providing feedback.

CLASS/FFT online system records. Through the online systems maintained by Teachstone (CLASS provider) and Teachscape (FFT provider), we gathered administrative records of classroom observations as well as observation scores. For each observation session, the system provided the names of the teacher and observer and indicated whether the observation and feedback sessions occurred.

VAL-ED online system records. The online system maintained by Discovery (VAL-ED provider) provided information about principal performance as well as administrative records regarding the number of teachers and district staff who were asked to complete the VAL-ED survey, the VAL-ED survey response rates, the dates when principals received the survey results, and the dates when principal feedback sessions occurred.

AIR online system records. AIR's online system reported value-added scores for all grades 4–8 mathematics and reading/English language arts (ELA) teachers in the treatment schools. In addition, the system reported school average value-added scores for each treatment school.

District interviews. Following semi structured protocols, trained interviewers conducted phone interviews in spring 2013 with officials in each school district who were responsible for teacher and principal performance management. These interviews, each lasting approximately 90 minutes, covered topics such as central office staff members' viewing of performance information provided by the intervention and their perception of the clarity and usefulness of the information. The interviews also collected information about the integration of the study's intervention with existing district processes and information about future plans for the districts' educator evaluation systems. In addition, the interviews gathered contextual information regarding the districts' human resources policies (i.e., business as usual), focusing on their teacher and principal evaluation system policies and the ways in which performance data were used.

Educator Performance

Data on measures of teacher classroom practice, student growth, and principal leadership were collected from the providers' online systems throughout the study year. These data were collected only for teachers and principals in the treatment schools.

Educators' Experiences

Data on educators' experiences with performance evaluation were collected through a teacher survey and a principal survey in spring 2013. The teacher survey was administered to all K–8 teachers of mathematics and reading/ELA in the 127 study schools to collect data about teachers' experiences with performance evaluation. The survey took about 30 minutes to complete. The overall response rate was 99.5 percent for all teachers surveyed and 99.3 percent for grades 4–8 teachers.

The principal survey was administered to the principal of each study school to collect data about principals' experiences with performance evaluation. The survey took about 30 minutes to complete. The overall response rate was 96.9 percent.

Participant Characteristics

To compare the characteristics of participants in the treatment and control groups, we collected data on school characteristics from the 2011–12 Common Core of Data and collected data on the characteristics of principals, teachers, and students in study schools from district administrative records in the summer and fall of 2012.

Appendix C. Technical Details About Reliability Estimation

In this appendix, we describe the methods used to estimate the reliability of educator performance measures discussed in the report. The appendix begins with an overview of how reliability was conceptualized for this study. We then describe the methods used to estimate reliability for different aspects of the study's performance measures:

- the teacher classroom practice measures;
- differences between the scores a teacher received for different dimensions of classroom practice;
- the student growth measure (i.e., teacher value-added scores);
- differences between the value-added subject scores a teacher received;
- the principal leadership measure; and
- differences between the scores a principal received for different dimensions of principal leadership.

We estimated the reliability of the educator performance measures to describe the extent to which the measures implemented for the intervention provide consistent information about educator performance (i.e., the extent to which the measures are an indicator of an educator's true performance). The reliability estimation methods differed across the measures based on the data available for each measure and the inferences we sought to make in the report. Each method has limitations, and the estimated reliabilities are specific to the study context. For example, the estimated reliabilities for the classroom practice measures may depend on how observers were trained, the number of observers and observations, and the sample of classrooms observed. Since such conditions can differ from study to study, it is important to examine reliability within the specific context of this study, rather than rely on reliabilities reported in other studies. Unless otherwise stated, the reported reliability estimates represent the reliability of "absolute" scores (i.e., the consistency of educators' performance on a fixed metric) rather than the reliability of "relative" scores (i.e., the consistency of educators' standing relative to other educators), the former of which provides a more conservative reliability estimate (Webb, Shavelson, and Haertel 2006). While reliabilities above .60 or .70 are generally considered acceptable in the educational research literature, the acceptable level of reliability of a measure depends on the intended use (e.g., staffing decisions, professional development decisions), which affects the costs of misclassifying educators based on their scores.

Overview of Reliability

Measures of teacher and principal performance, like any measure, are susceptible to measurement error, which can artificially inflate the amount of variation in the observed ratings and undermine the ratings' utility. Using a generalizability theory framework (Shavelson and Webb 1991), reliability can be defined based on how much variation in a measure's ratings is the result of "true" differences in subjects rather than measurement error. In general, if we know the magnitude of the measurement error from different sources, then we can determine a measure's true score variance (i.e., total observed variance minus error variance) and calculate the measure's reliability as: (true score variance) / (true score variance + error variance).

Measurement error can arise from different sources depending on the measurement design. For the measure of teacher classroom practice in this study, which was based on one observation from a school administrator and three from a study-hired observer during a school year, we are primarily concerned about measurement error arising from the following seven sources of error:

1. *Systematic differences across observers.* The extent to which teacher ratings differ across observers (also known as observer severity, e.g., some observers always give higher ratings than other observers)
2. *Systematic differences across occasions.* The extent to which teacher ratings differ from lesson to lesson and day to day (e.g., all teachers get higher ratings with some types of lessons than others or at a certain time of the year than at other times)
3. *Teacher-by-observer differences.* The extent to which observer judgment differs based on the type of teacher observed (e.g., some observers tend to give higher scores to female teachers than to male teachers)
4. *Teacher-by-occasion differences.* The extent to which the ratings on particular occasions differ based on the type of teacher (e.g., teachers happen to receive an abnormally high rating on a day when low-achieving and disruptive students were absent or some teachers perform better on Friday afternoons while other teachers perform worse)
5. *Observer-by-occasion differences.* The extent to which observer judgment differs based on the lesson or day observed (e.g., observers happen to give abnormally lower ratings when observing before lunch)
6. *Teacher-by-observer-by-occasion differences.* The extent to which ratings differ because of specific combinations of how teacher performance and observer judgment change from occasion to occasion (e.g., some observers give abnormally low ratings when observing male teachers on Mondays)
7. *Random error.* The extent to which ratings differ for unknown or idiosyncratic reasons

Similar sources of error exist for the measure of teacher contributions to student achievement growth (i.e., value added) and the measure of principal leadership. For the measure of teacher contributions to student achievement growth, value-added scores were based on the achievement test scores from a teacher's classes in the prior two years. Therefore, one can think of students as analogous to observers because each student test score is used to "rate" teacher performance and years as analogous to occasions because the context within which teacher performance is assessed changes from one year to the next. For the measure of principal leadership, VAL-ED scores were based on ratings from three types of "observers" (i.e., principals, principals' supervisors, and teachers) in two occasions (i.e., assessment window).

Estimating the Reliability of the Intervention's Measures of Teacher Classroom Practice

We estimated the reliability of the teacher classroom practice ratings as a measure of stable classroom practice quality over a year. While a teacher's actual classroom practice could improve during the course of the year in response to factors such as feedback and professional development, we estimated the reliability with which the observations captured a teacher's "persistent," or average practice, during the year. In this study, a teacher was never rated by two different observers on the same occasion, so we could not directly identify the sources of error outlined above. In particular, we could not distinguish observer-based sources of error from occasion-based sources of error because observers were confounded with occasions. We were, however, able to estimate the amount of error from combined sources involving observers and occasions when analyzing the variation in ratings over the four observation windows. We refer to reliability based on variation in ratings over the observation windows as *intertemporal reliability*, or the proportion of variation in the teacher ratings that reflects stable differences among teachers in their classroom practice over the year.

We estimated intertemporal reliability in two steps. In the first step, we estimated the amount of between-teacher (representing persistent differences in ratings between teachers) and within-teacher variation (error variance from sources involving raters and occasions and random errors) based on scores from the four observation windows. In the second step, we use estimates from the first step and a set of assumptions about observer-based error and occasion-based error to calculate plausible reliability estimates for the four-window average scores. The following paragraphs describe the approach in more detail.

For the first step, we used a two-level hierarchical linear model (ratings nested in teachers) to decompose the total variation in the scores from the four observation windows into between-teacher variation and within-teacher variation. In practice, teachers are typically compared with other teachers within the same district, so we included district fixed effects in the model. With district fixed effects, the variance estimates reflect within-district variation in teacher scores and average between-district differences do not influence the reliability estimates. The variance decomposition results for the overall score and dimension scores are presented in exhibit C.1 for CLASS and exhibit C.2 for FFT. The proportion of between-teacher variance represents the inter-temporal reliability of a score based on one observation and one rater:

$$\frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2}$$

where σ_b^2 is the estimated between-teacher variance and σ_w^2 is the estimated within-teacher variance.

For the second step, the intertemporal reliability of the four-window average score depends on how much of the within-teacher variance was due to observer-based sources of error versus occasion-based sources of error. Since teachers typically had two observers during the year (a school administrator and a study-hired observer), calculating the reliability of the four-window average score requires dividing observer-based sources of error by two and dividing occasion-based sources of error by four. The available data did not allow us to disentangle observer-based error from occasion-based error, so we calculated reliability under different assumptions about the proportion of within-teacher variance due to observer-based sources of error. In the right-side columns of exhibits C.1 and C.2, we report the four-window reliability estimates under the following alternative assumptions:

- Zero percent of the error variance was observer-based error and 100 percent was occasion-based error.
- Twenty-five percent of the error variance was observer-based error and 75 percent was occasion-based error.
- Fifty-five percent of the error variance was observer-based error and 50 percent was occasion-based error.
- Seventy-five percent of the error variance was observer-based error and 25 percent was occasion-based error.
- One hundred percent of the error variance was observer-based error and 0 percent was occasion-based error.

Under a given assumption, the four-window reliability estimate is based on the following equation:

$$\frac{\sigma_b^2}{\sigma_b^2 + \frac{\pi_o \sigma_w^2}{2} + \frac{(1-\pi_o)\sigma_w^2}{4}}$$

where σ_b^2 is the estimated between-teacher variance, σ_w^2 is the estimated within-teacher variance, and π_o is the assumed proportion of error variance due to observer-based error. The plausible estimates of the four-window reliability reported in chapter 2 do not include the estimates based on an assumption of zero observer-based error or zero occasion-based error because such extremes are unlikely.

The reliability estimates presented in exhibits C.1 and C.2 are generally consistent with the findings from other studies of the variation in classroom observation ratings. To compare our estimates with findings from other studies, we can focus on the percentage of within-teacher variation, or error variance, and the percentage of between-teacher variation, which represents the reliability for ratings based on a single occasion and a single observer. We estimated that the reliability for ratings based on a single occasion and observer (between-teacher variation) was .24 and .49 for CLASS and FFT, respectively. Other studies suggest that the reliabilities for specific CLASS domain scores are between .13 and .35 based on a single occasion and observer (Casabianca et al. 2013), and the reliability of FFT is between .27 and .45 (Ho and Kane 2013).[85]

[85] Since we could not distinguish between occasion-based and observer-based error, it is informative to consider what other studies found for the percent of variation due to occasions and observers. The MET project, for example, found that 6 percent to 13 percent of the variation in CLASS or FFT scores was a result of variation between observers and 7 percent to 27 percent was a result of variation between occasions (Ho and Kane 2013; Kane and Staiger 2012). A separate study of CLASS domain scores (Casabianca et al. 2013) found that observer variation

These low reliabilities for ratings based on a single occasion and a single observer are why it is generally recommended to conduct classroom observations over multiple occasions and use multiple observers, which increases reliability by “averaging over” errors associated with occasions and observers.

accounted for 5 percent to 30 percent of the total variation in domain scores and occasion variation accounted for 13 percent to 18 percent of the total variation.

Exhibit C.1. Estimated reliabilities for CLASS overall scores and dimension scores

	Variance estimate		Proportion of variance		Four-window average reliability estimate under different assumptions				
CLASS dimensions	Between teacher	Within teacher	Between teacher[a]	Within teacher	0% observer error	25% observer error	50% observer error	75% observer error	100% observer error
Overall score	0.15	0.48	.24	.76	.56	.50	.46	.42	.39
Domain: Emotional support									
Positive climate	0.16	0.83	.16	.84	.43	.38	.34	.30	.28
Teacher sensitivity	0.12	0.85	.12	.88	.36	.31	.27	.24	.22
Regard for student perspectives	0.29	0.98	.23	.77	.54	.49	.44	.40	.37
Domain: Classroom organization									
Behavior management	0.18	0.79	.19	.81	.48	.43	.38	.35	.32
Productivity	0.12	0.68	.15	.85	.41	.36	.32	.29	.26
Negative climate (reverse coded)	0.02	0.36	.05	.95	.16	.13	.11	.10	.09
Domain: Instructional support									
Instructional learning formats	0.19	0.81	.19	.81	.48	.42	.38	.34	.31
Content understanding	0.21	0.90	.19	.81	.48	.43	.38	.35	.32
Analysis and inquiry	0.36	1.37	.21	.79	.52	.46	.42	.38	.35
Quality of feedback	0.30	1.15	.21	.79	.51	.46	.41	.38	.34
Instructional dialogue	0.31	1.26	.20	.80	.50	.44	.40	.36	.33
Domain: Student engagement	0.19	0.71	.21	.79	.51	.46	.41	.38	.35

NOTE: Sample size = 313 teachers.

[a] The proportion of between-teacher variance is also the reliability for ratings based on a single occasions and a single observer.

SOURCE: Teachstone Online System.

Exhibit C.2. Estimated reliabilities for FFT overall scores and dimension scores

FFT dimensions	Variance estimate		Proportion of variance		Four-window average reliability estimate under different assumptions				
	Between teacher	Within teacher	Between teacher [a]	Within teacher	0% observer error	25% observer error	50% observer error	75% observer error	100% observer error
Overall score	0.07	0.07	.49	.51	.79	.75	.72	.69	.66
Domain 2: Classroom environment									
Creating an environment of respect and rapport	0.09	0.20	.31	.69	.64	.59	.54	.51	.47
Establishing a culture for learning	0.09	0.20	.31	.69	.64	.59	.54	.50	.47
Managing classroom procedures	0.08	0.17	.32	.68	.65	.60	.55	.52	.48
Managing student behavior	0.10	0.20	.33	.67	.66	.61	.56	.53	.49
Domain 3: Instruction									
Communicating with students	0.09	0.21	.31	.69	.64	.59	.55	.51	.47
Using questioning and discussion techniques	0.09	0.17	.34	.66	.67	.62	.58	.54	.51
Engaging students in learning	0.09	0.20	.30	.70	.64	.58	.54	.50	.47
Using assessment in instruction	0.06	0.21	.24	.76	.55	.50	.45	.41	.38

NOTE: We refer to the FFT “components” as “dimensions” for consistency of terminology throughout the report. Reliability estimates for two components, organizing physical space and demonstrating flexibility and responsiveness, were not reported because observers did not rate these two components in each observation window. Sample size = 222 teachers.

[a] The proportion of between-teacher variance is also the reliability for ratings based on a single occasions and a single observer.

SOURCE: Teachscape Online System.

Estimating the Reliability of Within-Teacher Differences Between Scores for Dimensions of Classroom Practice

The scores for specific dimensions of classroom practice can provide teachers with meaningful information about their relative performance in different dimensions of practice if differences between a teacher's scores reflect true differences in a teacher's performance and not just measurement error. To examine the extent to which differences between a teacher's scores reflect true differences in the teacher's performance in specific dimensions of classroom practice rather than idiosyncratic differences from various sources of error, we used analysis of variance (ANOVA) models and generalizability theory (Webb, Shavelson, and Haertel 2006) to estimate the reliability of difference scores. We specified fully crossed ANOVA models with scores based on teachers, dimension scores (CLASS dimensions or FFT components), and observation windows, where all facets were treated as random for the purposes of variance decomposition. With this model, the observed variance (σ_{obs}^2) is the sum of the following seven variance components:

$$\sigma_{obs}^2 = \sigma_t^2 + \sigma_w^2 + \sigma_d^2 + \sigma_{tw}^2 + \sigma_{td}^2 + \sigma_{wd}^2 + \sigma_{r,twd}^2$$

where each variance component is defined as follows:

- σ_t^2 = teacher variance
- σ_w^2 = window variance
- σ_d^2 = dimension variance
- σ_{tw}^2 = teacher-by-window variance
- σ_{td}^2 = teacher-by-dimension variance
- σ_{wd}^2 = window-by-dimension variance
- $\sigma_{r,twd}^2$ = residual variance, including teacher-by-window-by-dimension variance

With the estimated variance components, the reliability of difference scores based on a single observation is defined by the following equation:

$$\frac{\sigma_{td}^2}{\sigma_{td}^2 + \sigma_{wd}^2 + \sigma_{r,twd}^2}$$

where σ_{td}^2 is the estimated variance of the true difference scores, and $\sigma_{wd}^2 + \sigma_{r,twd}^2$ is the estimated error variance for the difference scores.

As with reliability estimation for the four-window average overall scores, the reliability of difference scores based on four-window average scores depends on the amount of variance due to observer-based sources of error and occasion-based sources of error. Since the available data do not allow us to distinguish these two sources of error from window-based variation, we calculated reliability under different assumptions about the proportion of window-based variation due to observer-based sources (π_o). Under a given assumption about π_o, the reliability of a difference score based on the four-window average scores can be estimated according to the following equation:

$$\frac{\sigma_{td}^2}{\sigma_{td}^2 + \frac{\pi_o \sigma_{wd}^2}{2} + \frac{(1-\pi_o)\sigma_{wd}^2}{4} + \frac{\pi_o \sigma_{r,twd}^2}{2} + \frac{(1-\pi_o)\sigma_{r,twd}^2}{4}}$$

The variance decomposition results and the reliability estimates for differences between dimension scores are presented in exhibit C.3 for CLASS and exhibit C.4 for FFT.

Exhibit C.3. Estimated variance components and reliabilities for dimension score differences

	CLASS		FFT	
Source of variance	**Estimated variance component**	**Proportion of total variance**	**Estimated variance component**	**Proportion of total variance**
teacher (*t*)	0.16	.11	0.05	.19
window (*w*)	0.02	.01	0.01	.02
dimension (*d*)	0.33	.22	0.01	.03
t x *w*	0.42	.28	0.04	.16
t x *d*	0.10	.07	0.01	.05
w x *d*	0.00	.00	0.00	.00
residual	0.44	.30	0.13	.54
Reliability estimates	CLASS		FFT	
Single-observation reliability	.19		.09	
Four-window average reliability estimate				
0% observer error	.48		.28	
25% observer error	.43		.23	
50% observer error	.38		.20	
75% observer error	.35		.18	
100% observer error	.32		.16	

NOTE: Sample size = 13,882 CLASS score (313 teachers × 4 windows × 12 dimensions) and 7,814 FFT scores (222 teachers × 4 windows × 10 components). Not all teachers had scores for all windows and all dimensions/components.

SOURCE: Teachstone Online System (CLASS) and Teachscape Online System (FFT).

Estimating the Reliability of the Intervention's Measure of Student Growth (i.e., Teacher Value-Added Scores)

We estimated the reliability of the teacher value-added scores as a measure of the stability of scores over the two years of student growth data that were used to calculate teacher value-added. While a teacher's true value-added could change over time, we estimated the reliability with which the value-added scores provided in the student growth reports captured a teacher's "persistent," or average practice, during the past two years. We refer to reliability based on variation in value-added scores across years as *intertemporal reliability*, or the proportion of variation in the teacher value-added scores that reflects stable differences among teachers in their performance over time.[86]

We estimated intertemporal reliability by decomposing the total variation in the scores from the two years into between-teacher variation (representing persistent differences in scores between teachers) and within-teacher variation (error variance from sources involving changes over each year and random errors). We used a two-level hierarchical linear model (annual scores nested in teachers) to estimate the within- and between-teacher variance. In practice, teachers are typically compared with other teachers within the same district, so we included district fixed effects in the model. With district fixed effects, the variance estimates reflect within-district variation in teacher scores, and average between-district differences are not included in the estimate of between-teacher variance.

The value-added scores were based on all grade 4-8 teachers in the districts, not just teachers in the study schools, and value-added scores based on less than ten students were suppressed in the student growth reports. Therefore, for the variance decomposition analysis, we used data for all grade 4-8 teachers with at least 10 students with data in each year. We ran separate models for reading/ELA and mathematics.

The variance decomposition results for each subject are presented in exhibit C.4. The proportion of between-teacher variance represents the intertemporal reliability of a value-added score based on one year of student growth data:

$$\frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2}$$

where σ_b^2 is the estimated between-teacher variance and σ_w^2 is the estimated within-teacher variance. The intertemporal reliability of a value-added score based on two years of student growth data is based on the following equation:

$$\frac{\sigma_b^2}{\sigma_b^2 + \frac{\sigma_w^2}{2}}$$

[86] The value-added scores provided to teachers were Empirical Bayes estimates. Because the Empirical Bayes estimates are shrunk toward the mean, the variance of the observed teacher scores is not the sum of the true variance plus error variance, and thus, the intertemporal reliability is not, strictly speaking, a reliability estimate. It can be interpreted as the proportional reduction in mean square error, which is analogous to reliability.

Exhibit C.4. Estimated reliabilities for value-added scores based on two years of student growth data

Subject	Variance estimate		Proportion of variance		Reliability based on two years
	Between teacher	Within teacher	Between teacher [a]	Within teacher	
Reading/English language arts	0.004	0.010	.28	.72	.44
Mathematics	0.022	0.021	.52	.48	.68

NOTE: Sample size = 977 teachers for reading/ELA; 964 teachers for mathematics.

[a] The proportion of between-teacher variance is also the reliability of the value-added scores if based on a single year of student growth data.

SOURCE: AIR value-added system.

Estimating the Reliability of Within-Teacher Value-Added Subject Differences

The value-added scores for specific subjects (i.e., mathematics and reading/ELA) can provide teachers with information about their relative performance in different subjects if differences between a teacher's subject-specific value-added scores reflect true differences in a teacher's performance and not just measurement error. To compare a teacher's performance in different subjects, first we had to determine a common metric with which we can compare a teacher's subject-specific value-added scores. We had two options for a common metric: (1) the teacher's value-added score in student test score standard deviation units or (2) the teacher's value-added percentile ranking. The two options could result in different conclusions about a teacher's relative performance in different subjects. For example, a teacher could have value-added scores of 0.3 in reading/ELA and 0.5 in mathematics, indicating the teacher did a better job raising student mathematics achievement than reading achievement. However, if both scores correspond to the 75th percentile rank, then one could conclude the teacher did equally well in both subjects compared with other teachers. For the purposes of estimating the reliability of within-teacher value-added subject differences, we used the value-added scores based on the student test score standard deviation unit, which is the raw metric used to estimate each teacher's value-added scores and corresponds to the value-added scores presented in chapter 3.

To examine the extent to which differences between a teacher's scores reflect true differences in the teacher's subject-specific performance rather than idiosyncratic differences from various sources of error, we used ANOVA models and generalizability theory (Webb, Shavelson, and Haertel 2006) to estimate the reliability of difference scores. We specified fully crossed ANOVA models with scores based on teachers, year of value-added score, and subject-specific scores, where all facets were treated as random for the purposes of variance decomposition. With this model, the observed variance (σ_{obs}^2) is the sum of the following seven variance components:

$$\sigma_{obs}^2 = \sigma_t^2 + \sigma_y^2 + \sigma_s^2 + \sigma_{ty}^2 + \sigma_{ts}^2 + \sigma_{ys}^2 + \sigma_{r,tys}^2$$

where each variance component is defined as follows:

- σ_t^2 = teacher variance
- σ_y^2 = year variance
- σ_s^2 = subject variance
- σ_{ty}^2 = teacher-by-year variance
- σ_{ts}^2 = teacher-by-subject variance
- σ_{ys}^2 = year-by-subject variance
- $\sigma_{r,tys}^2$ = residual variance, including teacher-by-year-by-subject variance

With the estimated variance components, the reliability of difference scores based on two years of value-added data is defined by the following equation:

$$\frac{\sigma_{ts}^2}{\sigma_{ts}^2 + \frac{\sigma_{ys}^2}{2} + \frac{\sigma_{r,tys}^2}{2}}$$

where σ_{ts}^2 is the estimated variance of the true difference scores, and $\sigma_{ys}^2 + \sigma_{r,tys}^2$ is the estimated error variance for the difference scores.

The variance decomposition results and the reliability estimates for differences between subject value-added scores are presented in exhibit C.5. The analysis was restricted to teachers with at least 10 students included in the value-added estimates for mathematics and reading/ELA in the two prior years. Restricting the analysis to value-added scores based on at least 10 students minimizes the extent to which these reliability estimates are driven by abnormal fluctuations in value-added scores due to small student sample sizes.

Exhibit C.5. Estimated variance components and reliability for subject-specific value-added score differences

Source of variance	Estimated variance component	Proportion of total variance
teacher (*t*)	0.01	.27
year (*y*)	0.00	.00
subject (*s*)	0.00	.00
t × *y*	0.01	.20
t × *s*	0.01	.18
y × *s*	0.00	.00
residual	0.01	.34
Reliability estimate	**.52**	

NOTE: Sample size = 2,772 value-added scores (693 teachers × 2 years × 2 subjects). The analysis included all teachers in the study districts with value-added scores based on at least 10 students in each year and subject.
SOURCE: AIR value-added system.

Estimating the Reliability of the Intervention’s Measure of Principal Leadership

We estimated the reliability of the principal leadership ratings as a measure of leadership quality within each assessment window (fall and spring). Since principals receive ratings from each of the three respondent groups, we estimated the reliability with which scores from the three groups captured a principal’s average leadership quality in the fall and spring. We refer to reliability based on variation in ratings between the respondent groups as *inter-rater reliability*, or the proportion of variation in the principal ratings that reflects respondent group agreement on each principal’s leadership quality. We did not examine the reliability of the principal leadership scores between the two assessment windows (i.e., intertemporal reliability) because the principal leadership reports and feedback emphasized how the principal did in each assessment window, and how the different respondent groups rated the principal in that window.

To estimate inter-rater reliability, we used a two-level hierarchical linear model (ratings nested in principals) to decompose the total variation in the scores from the three respondent groups into between-principal variation (representing consistent differences in ratings between principals) and within-principal variation (error variance from sources involving raters and random errors). In practice, principals are typically compared with other principals within the same district, so we included district fixed effects in the model. With district fixed effects, the variance estimates reflect within-district variation in principal scores and average between-district differences do not influence the reliability estimates. The variance decomposition results for the overall score and the dimension scores are presented in exhibit C.6 for fall and exhibit C.7 for spring. The proportion of within-principal variance represents the reliability of a score based on one respondent group. The inter-rater reliability for the score averaged across the three respondent groups is the reliability estimate presented in the last column of each exhibit and is based on the following equation:

$$\frac{\sigma_b^2}{\sigma_b^2 + \frac{\sigma_w^2}{3}}$$

where σ_b^2 is the estimated between-principal variance and σ_w^2 is the estimated within-principal variance.

Exhibit C.6. Estimated reliabilities for VAL-ED overall scores and dimension scores in the fall

VAL-ED Dimension	Variance estimate		Proportion of variance		Reliability estimate
	Between principal	Within principal	Between principal	Within principal	
Overall score	0.02	0.21	.07	.93	.19
Core components					
High standards for student learning	0.04	0.23	.14	.86	.32
Quality instruction	0.02	0.24	.08	.92	.21
Culture of learning and professional behavior	0.02	0.24	.09	.91	.24
Connections to external communities	0.02	0.25	.06	.94	.17
Performance accountability	0.02	0.28	.07	.93	.18
Rigorous curriculum	0.01	0.25	.03	.97	.08
Key processes					
Planning	0.02	0.23	.09	.91	.23
Implementing	0.02	0.21	.07	.93	.19
Supporting	0.01	0.22	.06	.94	.15
Advocating	0.02	0.22	.07	.93	.19
Communicating	0.01	0.27	.02	.98	.07
Monitoring	0.03	0.26	.11	.89	.28

NOTE: Sample size = 63 principals.
SOURCE: Fall 2012 VAL-ED Surveys.

Exhibit C.7. Estimated reliabilities for VAL-ED overall scores and dimension scores in the spring

VAL-ED dimension	Variance estimate		Proportion of variance		Reliability estimate
	Between principal	Within principal	Between principal	Within principal	
Overall score	0.06	0.18	.26	.74	.51
Core components					
High standards for student learning	0.09	0.17	.35	.65	.62
Quality instruction	0.06	0.20	.23	.77	.47
Culture of learning and professional behavior	0.07	0.22	.23	.77	.48
Connections to external communities	0.04	0.22	.14	.86	.33
Performance accountability	0.09	0.23	.29	.71	.55
Rigorous curriculum	0.06	0.20	.22	.78	.46
Key processes					
Planning	0.05	0.19	.22	.78	.46
Implementing	0.06	0.18	.25	.75	.50
Supporting	0.08	0.18	.30	.70	.56
Advocating	0.07	0.19	.26	.74	.51
Communicating	0.05	0.23	.19	.81	.41
Monitoring	0.07	0.20	.24	.76	.49

NOTE: Sample size = 63 principals.
SOURCE: Spring 2013 VAL-ED Surveys.

Estimating the Reliability of Within-Principal Differences Between Scores for Dimensions of Principal Leadership

The scores for specific dimensions of principal leadership can provide principals with information about their relative performance in different dimensions of leadership if differences between a principal's scores reflect true differences in a principal's performance and not just measurement error. For VAL-ED, dimensions of principal leadership are assessed in two interrelated ways: based on six core components and based on six key processes. Since the core components and key processes share assessment items, we conducted separate analyses for differences among the core components and differences among the key processes. To examine the extent to which differences between a principal's dimension scores reflect true differences in the principal's performance in specific dimensions of leadership rather than idiosyncratic differences from various sources of error, we used ANOVA models and generalizability theory (Webb, Shavelson, and Haertel 2006) to estimate the reliability of difference scores. We specified fully crossed ANOVA models with scores based on principals, dimension scores (core components or key processes), and respondent group (rater), where all facets were treated as random for the purposes of variance decomposition. With this model, the observed variance (σ_{obs}^2) is the sum of the following seven variance components:

$$\sigma_{obs}^2 = \sigma_p^2 + \sigma_r^2 + \sigma_d^2 + \sigma_{pr}^2 + \sigma_{pd}^2 + \sigma_{rd}^2 + \sigma_{e,prd}^2$$

where each variance component is defined as follows:

- σ_p^2 = principal variance
- σ_r^2 = rater variance
- σ_d^2 = dimension variance
- σ_{pr}^2 = principal-by-rater variance
- σ_{pd}^2 = principal-by-dimension variance
- σ_{rd}^2 = rater-by-dimension variance
- $\sigma_{e,prd}^2$ = residual variance, including principal-by-rater-by-dimension variance

With the estimated variance components, the reliability of difference scores based on average scores across the three respondent groups is defined by the following equation:

$$\frac{\sigma_{pd}^2}{\sigma_{pd}^2 + \frac{\sigma_{rd}^2}{3} + \frac{\sigma_{e,prd}^2}{3}}$$

where σ_{pd}^2 is the estimated variance of the true difference scores and $\frac{\sigma_{rd}^2}{3} + \frac{\sigma_{e,prd}^2}{3}$ is the estimated error variance for the difference scores averaged over the three respondent groups. The variance decomposition results and the reliability estimates for differences between scores are presented in exhibit C.8 for the fall wave and exhibit C.9 for the spring wave. We conducted separate analyses for the core components and key processes.

Exhibit C.8. Estimated variance components and reliabilities for VAL-ED dimension score differences: fall wave

	Core components		Key processes	
Source of variance	Estimated variance component	Proportion of total variance	Estimated variance component	Proportion of total variance
Principal (*p*)	0.03	.11	0.03	.11
Respondent group (*r*)	0.00	.00	0.00	.01
Dimension (*d*)	0.01	.03	0.00	.00
p × *r*	0.21	.70	0.21	.78
p × *d*	0.01	.03	0.00	.01
r × *d*	0.00	.01	0.00	.00
Residual	0.04	.13	0.02	.09
Reliability estimate	**.36**		**.29**	

NOTE: Sample size = 1,132 core component scores and 1,133 key process scores (63 principals × 3 respondent groups × 6 dimensions). Not all principals had scores from all respondent groups and all dimensions.

SOURCE: Fall 2012 VAL-ED Surveys.

Exhibit C.9. Estimated variance components and reliabilities for VAL-ED dimension score differences: spring wave

	Core components		Key processes	
Source of variance	Estimated variance component	Proportion of total variance	Estimated variance component	Proportion of total variance
Principal (*p*)	0.07	.22	0.07	.26
Respondent group (*r*)	0.02	.05	0.02	.06
Dimension (*d*)	0.01	.04	0.00	.01
p × *r*	0.16	.54	0.16	.60
p × *d*	0.01	.04	0.00	.01
r × *d*	0.00	.01	0.00	.00
Residual	0.03	.11	0.02	.07
Reliability estimate	**.50**		**.20**	

NOTE: Sample size = 1,133 core component scores and 1,133 key process scores (63 principals × 3 respondent groups × 6 dimensions). Not all principals had scores from all respondent groups and all dimensions.

SOURCE: Spring 2013 VAL-ED Surveys.

Appendix D. Technical Details About Analyses Assessing Treatment-Control Differences in Educators' Experiences

This appendix includes the technical details for statistical analyses examining treatment-control differences in educators' experiences during the first year of the study.

To assess whether the intervention led to differences in educators' experiences with performance evaluation (i.e., service contrast), we compared the survey responses of educators in the treatment schools with the responses of educators in the control schools. Our analytic approach differs for binary survey measures (e.g., whether a teacher received feedback based on observations) and continuous survey measures (e.g., the number of instances of feedback received), as described separately next.

Analyses of binary measures. For binary measures of educators' experiences, we examined the treatment-control differences using a principal-level linear probability model for principal survey measures and a two-level linear probability model for teacher survey measures, as specified next.[87]

Linear probability model to estimate treatment-control differences in binary principal survey measures:

$$Y_k = \sum_{b=1}^{37} \gamma_{0b} B_{bk} + \sum_{d=1}^{8} \gamma_{1d} (T * D_d)_k + u_k$$

where

- Y_k is the response of principal *k* to a given binary survey measure;
- B_{bk} , *b* = 1–37, is a set of dummy indicators for the 37 random assignment blocks;
- $(T * D_d)_k$, *d* = 1–8, is a set of treatment-by-district interactions; and
- u_k is a random error associated with principal *k*.

The estimate of primary interest from the above model is γ_{1d} , *d* = 1–8, which represents the treatment-control difference in the principal survey measure in each of the eight study districts.

[87] We decided to use a linear probability mode for binary survey measures because a logit model would encounter the quasi-complete separation problem (Albert and Anderson 1984; Allison 2008) for some of the binary measures, which occurs if 100 percent of the treatment principals and teachers or 100 percent of the control principals and teachers within some districts experienced the outcome. For such districts, the district-specific treatment effects cannot be estimated because the maximum likelihood estimates do not exist.

These eight district-specific differences were then combined into a weighted average difference, with each district weighted by the number of treatment schools in the district.

Two-level linear probability model to estimate treatment-control differences in binary teacher survey measures:

Level 1 (teachers)

$$Y_{jk} = \beta_{0k} + r_{jk}$$

Level 2 (schools)

$$\beta_{0k} = \sum_{b=1}^{37} \gamma_{00b} B_{bk} + \sum_{d=1}^{8} \gamma_{01d} (T * D_d)_k + u_{0k}$$

The model to estimate the treatment-control differences in binary teacher survey measures is similar to the model for binary principal survey measures, with the only difference being that the model for teacher survey measures is specified as a two-level model to account for the clustering of teachers within schools. The estimate of primary interest from the above model is γ_{01d} , d = 1–8, which represents the treatment-control difference in the teacher survey measure in each of the 8 study districts. These eight district-specific differences were then combined into a weighted average difference, with each district weighted by the number of treatment schools in the district.

Analyses of Continuous Measures. For continuous survey measures of principals' and teachers' experiences with performance evaluation, we estimated the treatment-control differences by comparing the median survey responses from the two study groups using nonparametric analyses because many of the survey-based continuous variables do not meet the distributional assumptions for parametric analysis. Specifically, all the survey-based continuous variables analyzed for this report are either measures of counts (e.g., number of instances of feedback) or measures of duration (e.g., length of oral feedback).

Many of these measures are not normally distributed due to the presence of outliers or an excess of zeros, which make normal theory inference statistics (such as the *p* value) based on standard parametric methods invalid. Moreover, while the average difference between the treatment and control groups is often the most informative statistic, the presence of outliers and the overabundance of zeros make it a potentially misleading description of the typical difference between treatment and control educators.

Nonparametric models are particularly well suited to data that do not meet the distributional assumptions underlying standard parametric analysis because they are "distribution free." The specific nonparametric model we used to analyze the continuous survey measures is the aligned rank sum test (Hodges and Lehmann 1962). The test is a regression-adjusted version of the Wilcoxon rank sum test, also called the Mann-Whitney U test, which is the most commonly used nonparametric test. The aligned rank sum test estimates a median treatment effect with or without covariate adjustment while making no distributional assumptions about the error terms. The test also has been shown to have a considerable efficiency advantage, relative to a normal theory estimator, when the residuals are not normally distributed (Blair and Higgins 1980; Kitchen 2009). For the analyses estimating treatment-control differences in survey measures of educators' experiences, the aligned rank sum test accounted for block fixed effects but not other covariates and was implemented in R.

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

Appendix E. Supplemental Findings About the Implementation of the Intervention's Measures of Classroom Practice

Exhibit E.1. Percentage of observers who agreed somewhat or strongly with each statement about the fairness and validity of CLASS or FFT

Statement about CLASS/FFT	CLASS		FFT	
	Principals	Study-hired observers	Principals	Study-hired observers
The rating system does a good job distinguishing effective from ineffective teaching.	≥ 89.0†	85.7	≥ 89.0†	90.3
The rating system is fair to all teachers, regardless of their personal characteristics or those of the students they teach.	≥ 89.0†	≥ 89.0†	≥ 89.0†	71.0
The rating system accurately reflects the quality of an individual's teaching.	≥ 89.0†	≥ 89.0†	90.0	≥ 89.0†

NOTE: Sample size = 31 principals and 28 study-hired observers for CLASS and 30 principals and 30–31 study-hired observers for FFT.

† Exact percentages are suppressed due to small number of principals or study-hired observers who disagreed with the statement.

SOURCE: Spring 2013 Principal and Study-Hired Observer Surveys.

Exhibit E.2. Percentage of K–3 teachers who received zero, one, or two study observations and feedback sessions in CLASS and FFT districts

Districts	Number of observations			Number of feedback sessions		
	0	1	2	0	1	2
CLASS	0.0	4.3	95.7	18.9	4.5	76.6
FFT	0.0	2.9	97.1	0.0	3.3	96.7
All	**0.0**	**3.7**	**96.3**	**10.9**	**4.0**	**85.1**

NOTE: Sample size = 24 schools and 376 teachers in CLASS districts; 25 schools and 276 teachers in FFT districts.

SOURCE: CLASS and FFT Provider Systems.

Exhibit E.3. Descriptive statistics for four-window average CLASS observation scores, by domain and dimension

CLASS domains and dimensions	*N*	Mean	Standard deviation	Minimum	25th percentile	75th percentile	Maximum
Overall score	313	5.72	0.56	3.63	5.33	6.10	6.99
Domain: Emotional support	313	5.75	0.63	3.79	5.33	6.17	7.00
Positive climate	313	6.04	0.64	3.75	5.67	6.50	7.00
Teacher sensitivity	313	5.91	0.62	4.00	5.50	6.38	7.00
Regard for student perspectives	313	5.30	0.84	2.50	4.75	5.88	7.00
Domain: Classroom organization	313	6.42	0.44	4.78	6.21	6.75	7.00
Behavior management	313	6.23	0.64	3.75	5.88	6.75	7.00
Productivity	313	6.22	0.58	4.00	5.88	6.67	7.00
Negative climate (reverse coded)	313	6.81	0.35	4.75	6.75	7.00	7.00
Domain: Instructional support	313	5.20	0.77	2.60	4.70	5.78	7.00
Instructional learning formats	313	5.68	0.66	2.67	5.25	6.13	7.00
Content understanding	313	5.35	0.77	2.33	4.75	5.88	7.00
Analysis and inquiry	313	4.72	1.05	1.83	4.00	5.50	7.00
Quality of feedback	313	5.24	0.83	2.63	4.67	5.83	7.00
Instructional dialogue	313	5.01	0.93	2.75	4.33	5.75	7.00
Domain: Student engagement	313	6.10	0.62	3.50	5.75	6.50	7.00

SOURCE: Teachstone Online System.

Exhibit E.4. Descriptive statistics for four-window average FFT observation scores, by dimension

FFT dimensions	*N*	Mean	Standard deviation	Minimum	25th percentile	75th percentile	Maximum
Overall score	222	3.06	0.29	1.97	2.91	3.25	3.97
Domain 2: Classroom environment							
Creating an environment of respect and rapport	222	3.20	0.38	2.00	3.00	3.50	4.00
Establishing a culture for learning	222	3.04	0.37	1.75	2.75	3.25	4.00
Managing classroom procedures	222	3.05	0.35	2.00	3.00	3.25	4.00
Managing student behavior	222	3.07	0.39	1.75	3.00	3.25	4.00
Organizing physical space	205	3.07	0.32	2.00	3.00	3.00	4.00
Domain 3: Instruction							
Communicating with students	222	3.21	0.38	2.25	3.00	3.50	4.00
Using questioning and discussion techniques	222	2.92	0.37	1.50	2.75	3.00	3.75
Engaging students in learning	222	3.00	0.37	1.75	2.75	3.25	4.00
Using assessment in instruction	221	2.98	0.36	1.75	2.75	3.25	4.00
Demonstrating flexibility and responsiveness	196	3.01	0.33	1.50	3.00	3.00	4.00

SOURCE: Teachscape Online System.

Exhibit E.5. Descriptive statistics for four-window average CLASS observation scores, by teacher characteristics and domain

Teacher characteristics		Overall score		Emotional support		Classroom organization		Instructional support		Student engagement	
	N	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
All teachers	313	5.72	0.56	5.75	0.63	6.42	0.44	5.20	0.77	6.10	0.62
Grade level											
4 and 5	182	5.85	0.53	5.87	0.58	6.49	0.39	5.38	0.77	6.26	0.56
6–8	131	5.53	0.56	5.59	0.65	6.31	0.48	4.95	0.70	5.88	0.64
Subject taught											
General	179	5.86	0.54	5.89	0.59	6.52	0.40	5.37	0.77	6.26	0.57
Mathematics	69	5.43	0.61	5.40	0.69	6.23	0.50	4.90	0.75	5.75	0.71
Reading/ELA	65	5.63	0.46	5.74	0.51	6.34	0.42	5.05	0.65	6.04	0.52
Years of experience											
0–3	35	5.50	0.66	5.53	0.68	6.18	0.56	5.03	0.89	5.77	0.72
4–10	104	5.79	0.48	5.82	0.55	6.45	0.34	5.30	0.68	6.18	0.58
11–20	84	5.72	0.53	5.76	0.56	6.43	0.49	5.20	0.72	6.09	0.58
20+	83	5.74	0.60	5.75	0.72	6.50	0.38	5.18	0.81	6.18	0.60

SOURCE: Teachstone Online System.

Exhibit E.6. Descriptive statistics for four-window average FFT observation scores, by teacher characteristics and domain

Teacher characteristics	Overall score			Classroom environment		Instruction	
	N	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
All teachers	222	3.06	0.29	3.09	0.31	3.03	0.31
Grade level							
4 and 5	119	3.08	0.26	3.11	0.26	3.06	0.26
6–8	103	3.03	0.33	3.07	0.35	2.99	0.35
Subject taught							
General	158	3.07	0.27	3.11	0.27	3.05	0.27
Mathematics	31	3.05	0.35	3.09	0.41	3.00	0.41
Reading/ELA	32	2.99	0.37	3.03	0.37	2.94	0.37
Years of experience							
0–3	33	2.98	0.38	3.04	0.39	2.92	0.39
4–10	81	3.10	0.30	3.12	0.31	3.08	0.31
11–20	54	3.08	0.24	3.12	0.26	3.04	0.26
20+	49	3.02	0.30	3.06	0.30	2.99	0.30

SOURCE: Teachscape Online System.

Exhibit E.7. Descriptive statistics for average CLASS observation scores, by observer type

CLASS domains and dimensions	Score from study-hired observers			Score from school administrators			Correlation coefficient[a]
	N	Mean	Standard deviation	*N*	Mean	Standard deviation	
Overall score	294	5.80	0.58	245	5.54	0.82	.40
Domain: Emotional support	294	5.87	0.63	245	5.51	0.93	.36
Positive climate	294	6.16	0.66	245	5.82	1.05	.25
Teacher sensitivity	294	6.02	0.66	245	5.66	0.96	.24
Regard for student perspectives	294	5.43	0.87	245	5.04	1.11	.40
Domain: Classroom organization	294	6.49	0.47	245	6.30	0.66	.34
Behavior management	294	6.31	0.69	245	6.13	0.93	.29
Productivity	294	6.34	0.57	245	5.99	0.98	.35
Negative climate (reverse coded)	294	6.82	0.38	245	6.79	0.62	.05
Domain: Instructional support	294	5.27	0.82	245	5.00	1.04	.40
Instructional learning formats	294	5.82	0.70	245	5.39	0.97	.29
Content understanding	294	5.41	0.85	245	5.10	1.09	.38
Analysis and inquiry	294	4.84	1.09	244	4.49	1.40	.36
Quality of feedback	294	5.25	0.93	245	5.18	1.15	.35
Instructional dialogue	294	5.05	1.04	245	4.85	1.22	.35
Domain: Student engagement	294	6.14	0.67	244	6.01	0.99	.31

NOTE: In cases where a teacher had more than one score from a school administrator, the average score was used. The mean difference between the overall score from study-hired observers and the overall score from school administrators was statistically significant (p < .05).

[a] Correlation coefficients are based on correlations between teachers' mean score from study-hired observers (averaged across multiple observations) and score from the school administrator.

SOURCE: Teachstone Online System.

Exhibit E.8. Descriptive statistics for average FFT observation scores, by observer type

FFT dimensions	Score from study-hired observers			Score from school administrator			Correlation coefficient[a]
	N	Mean	Standard deviation	*N*	Mean	Standard deviation	
Overall score	222	3.07	0.31	221	3.04	0.38	.56
Domain 2: Classroom environment							
Creating an environment of respect and rapport	222	3.21	0.42	218	3.20	0.51	.32
Establishing a culture for learning	222	3.05	0.43	220	3.01	0.48	.27
Managing classroom procedures	222	3.05	0.37	219	3.06	0.54	.40
Managing student behavior	222	3.07	0.41	220	3.10	0.56	.41
Organizing physical space	187	3.10	0.41	122	3.02	0.37	.14
Domain 3: Instruction							
Communicating with students	222	3.22	0.42	221	3.20	0.54	.40
Using questioning and discussion techniques	222	2.94	0.37	218	2.89	0.57	.45
Engaging students in learning	222	3.02	0.41	221	2.97	0.49	.38
Using assessment in instruction	220	3.01	0.38	214	2.93	0.55	.28
Demonstrating flexibility and responsiveness	171	3.05	0.34	117	2.92	0.44	.03

NOTE: In cases where a teacher had more than one score from a school administrator, the average score was used. The mean difference between the overall score from study-hired observers and the overall score from school administrators was not statistically significant (*p* = .111).

[a] Correlation coefficients are based on correlations between teachers' mean score from study-hired observers (averaged across multiple observations) and score from the school administrator.

SOURCE: Teachscape Online System.

Exhibit E.9. Descriptive statistics for four-window average CLASS observation scores for K–3 teachers, by domain and dimension

CLASS domains and dimensions	*N*	Mean	Standard deviation	Minimum	25th percentile	75th percentile	Maximum
Overall score	376	5.70	0.64	3.20	5.29	6.15	6.95
Domain: Emotional support	376	6.12	0.52	3.75	5.88	6.44	7.00
Positive climate	376	6.24	0.69	3.00	5.88	6.75	7.00
Negative climate (reverse coded)	376	6.87	0.31	4.00	7.00	7.00	7.00
Teacher sensitivity	376	5.95	0.73	3.00	5.50	6.50	7.00
Regard for student perspectives	376	5.43	0.90	2.75	4.75	6.00	7.00
Domain: Classroom organization	376	5.89	0.72	2.33	5.54	6.42	7.00
Behavior management	376	6.05	0.86	2.50	5.75	6.75	7.00
Productivity	376	6.02	0.81	2.00	5.50	6.50	7.00
Instructional learning formats	376	5.60	0.76	2.00	5.25	6.00	7.00
Domain: Instructional support	376	4.94	1.04	2.25	4.17	5.75	7.00
Content development	376	4.82	1.17	2.00	3.75	5.75	7.00
Quality of feedback	376	5.10	1.05	2.00	4.38	6.00	7.00
Language modeling	376	4.91	1.06	2.25	4.25	5.75	7.00

SOURCE: Teachstone Online System.

Exhibit E.10. Descriptive statistics for four-window average FFT observation scores for K–3 teachers, by dimension

FFT dimensions	*N*	Mean	Standard deviation	Minimum	25th percentile	75th percentile	Maximum
Overall score	276	3.02	0.25	2.06	2.90	3.16	4.00
Domain 2: Classroom environment							
Creating an environment of respect and rapport	275	3.19	0.42	2.00	3.00	3.50	4.00
Establishing a culture for learning	276	2.99	0.31	2.00	3.00	3.00	4.00
Managing classroom procedures	276	3.01	0.39	2.00	3.00	3.00	4.00
Managing student behavior	276	3.04	0.40	1.00	3.00	3.00	4.00
Organizing physical space	220	3.06	0.33	2.00	3.00	3.00	4.00
Domain 3: Instruction							
Communicating with students	276	3.14	0.35	2.00	3.00	3.50	4.00
Using questioning and discussion techniques	274	2.88	0.35	1.50	2.50	3.00	4.00
Engaging students in learning	276	3.01	0.35	2.00	3.00	3.00	4.00
Using assessment in instruction	276	2.93	0.39	1.50	2.75	3.00	4.00
Demonstrating flexibility and responsiveness	192	3.02	0.39	1.00	3.00	3.00	4.00

SOURCE: Teachscape Online System.

Exhibit E.11. Correlations between the window-specific overall scores

	CLASS		FFT	
Pairwise correlations	*N*	Correlation coefficient	*N*	Correlation coefficient
Window 1 and Window 2	260	.21	215	.54
Window 1 and Window 3	259	.25	214	.53
Window 1 and Window 4	228	.21	211	.52
Window 2 and Window 3	304	.39	217	.59
Window 2 and Window 4	273	.15	214	.54
Window 3 and Window 4	278	.45	216	.50

SOURCE: Teachstone Online System (CLASS) and Teachscape Online System (FFT).

Exhibit E.12. Correlations between the four-window average CLASS dimension scores

Domain	Dimension	E1	E2	E3	C1	C2	C3	I1	I2	I3	I4	I5	S1
Emotional Support	(E1) Positive climate	1.00											
	(E2) Teacher sensitivity	.85	1.00										
	(E3) Regard for student perspectives	.64	.64	1.00									
Class. Org.	(C1) Behavior management	.63	.65	.39	1.00								
	(C2) Productivity	.60	.67	.44	.73	1.00							
	(C3) Negative climate (reverse coded)	.39	.34	.12	.46	.36	1.00						
Instr. Support	(I1) Instructional learning formats	.65	.68	.69	.56	.66	.16	1.00					
	(I2) Content understanding	.47	.53	.70	.43	.45	.05	.73	1.00				
	(I3) Analysis and inquiry	.49	.51	.79	.32	.43	.00	.68	.82	1.00			
	(I4) Quality of feedback	.61	.65	.71	.49	.50	.11	.72	.80	.79	1.00		
	(I5) Instructional dialogue	.51	.54	.76	.35	.38	.00	.69	.81	.82	.85	1.00	
Std. Eng.	(S1) Student engagement	.66	.70	.59	.74	.73	.27	.74	.64	.57	.70	.64	1.00

NOTE: Sample sizes = 313 teachers. Shaded cells represent correlations between dimensions within the same domain.

SOURCE: Teachstone Online System.

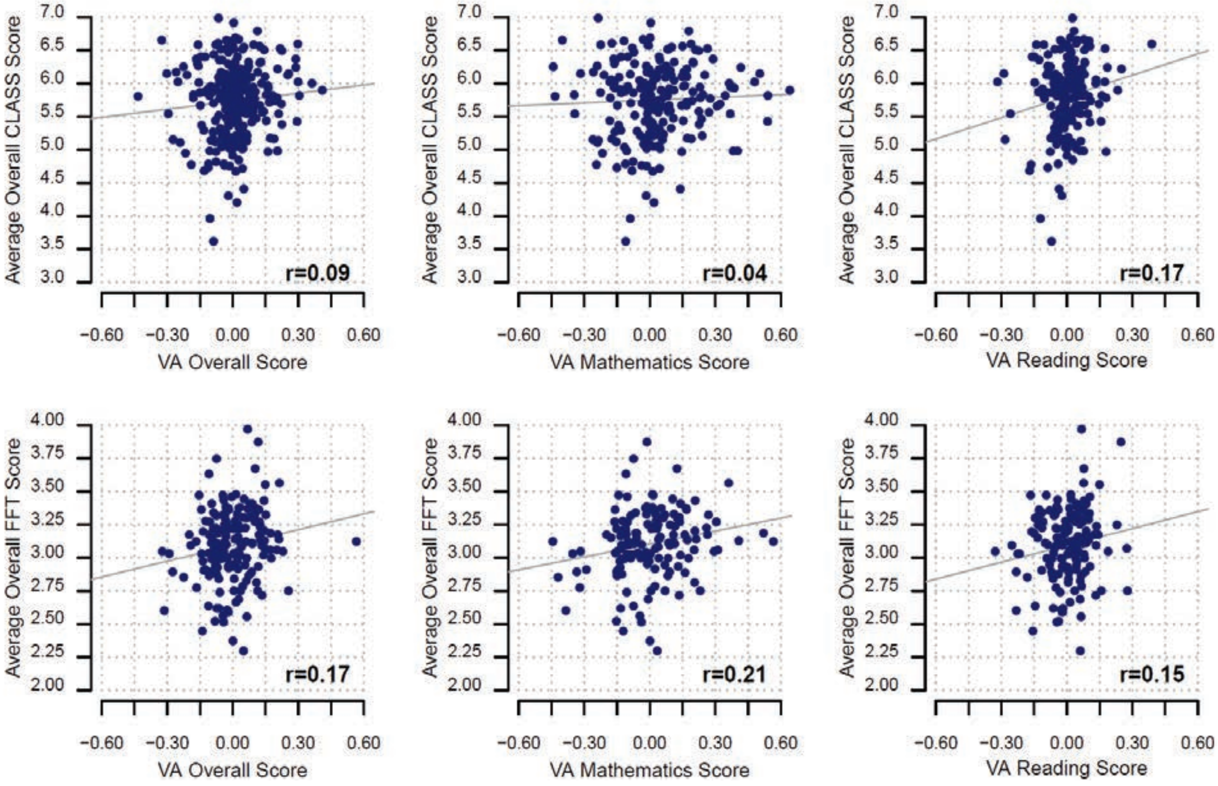
Exhibit E.13. Correlations between the four-window average FFT dimension scores

Domain	Dimension	B1	B2	B3	B4	B5	C1	C2	C3	C4	C5
Class. Env.	(B1) Creating an environment of respect and rapport	1.00									
	(B2) Establishing a culture for learning	.68	1.00								
	(B3) Managing classroom procedures	.64	.69	1.00							
	(B4) Managing student behavior	.71	.67	.72	1.00						
	(B5) Organizing physical space	.32	.39	.36	.29	1.00					
Instruct.	(C1) Communicating with students	.68	.68	.60	.63	.38	1.00				
	(C2) Using questioning and discussion techniques	.58	.69	.58	.58	.40	.67	1.00			
	(C3) Engaging students in learning	.64	.75	.66	.62	.35	.69	.75	1.00		
	(C4) Using assessment in instruction	.53	.64	.55	.54	.44	.63	.65	.66	1.00	
	(C5) Demonstrating flexibility and responsiveness	.36	.44	.35	.36	.33	.43	.47	.45	.51	1.00

NOTE: Pairwise correlations are reported, with pairwise sample sizes ranging from 195 to 222 teachers. Shaded cells represent correlations between dimensions within the same domain.

SOURCE: Teachscape Online System

Exhibit E.14. Relationships between teacher prior-year value-added scores and classroom observation four-window average overall scores



NOTE: The grey line represents the regression line for each relationship.

SOURCE: Teachstone Online System (CLASS), Teachscape Online System (FFT), and Student Growth Reporting System.

Exhibit E.15. Pairwise correlations between classroom observation overall scores and current-year value-added scores

	Overall[a]		Mathematics		ELA/Reading	
	N	Correlation coefficient	*N*	Correlation coefficient	*N*	Correlation coefficient
CLASS						
Four-window average	306	.18*	208	.20*	197	.10
Window 1	255	.16*	172	.20*	158	.04
Window 2	300	.11	204	.10	192	.03
Window 3	303	.16*	206	.17*	195	.18*
Window 4	273	.07	193	.10	177	.06
FFT						
Four-window average	214	.24*	177	.28*	179	.23*
Window 1	209	.19*	172	.21*	176	.18*
Window 2	211	.19*	174	.22*	177	.18*
Window 3	213	.21*	176	.21*	178	.26*
Window 4	210	.19*	175	.24*	176	.13

Exhibit Reads: The correlation between the four-window average CLASS overall scores and teachers' current year overall value-added scores was 0.18 based on 306 treatment teachers in CLASS districts.

NOTE: [a]The overall value-added score for a teacher with value-added scores in both mathematics and reading/ELA is a precision-weighted average of the value-added scores in both subjects. The overall value-added score is the same as the subject-specific value-added score for teachers with a value-added score in only one subject.

Two-tailed statistical significance at the $p < .05$ level is indicated by an asterisk (*).

SOURCE: Teachstone Online System (CLASS), Teachscape Online System (FFT), and Student Growth Reporting System.

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

Appendix F. Technical Details About the Estimation of Value-Added Scores

In this appendix, we describe technical details about the estimation of value-added scores provided to treatment teachers as part of the intervention. We first present the general specification of the value-added model, and then describe the covariates used in the model, which vary by district. In the last section, we explain how we calculated the overall value-added score for each teacher, school value-added scores, and district value-added scores based on the teacher-, subject-, grade-, and year-specific scores generated by the value-added model.

General Model Specification

The value-added model used for the study's intervention is a covariate adjustment model that includes the test scores for two prior years (where available), along with a set of measures of student characteristics (selected by districts), as predictor variables of current test scores, with students linked to specific teachers. Because there was a relatively small number of teachers per grade and subject in most of the study districts, no school effects were included in the model; that is, all between-teacher variance in students' achievement (controlling for measured covariates) was attributed to teachers, with no common variance attributed to their schools. The model uses an errors-in-variables regression approach to account for the measurement error in both prior and current test scores.[88]

The value-added model was estimated separately by grade, subject, and district, with the following general form:

$$\mathrm{y}_{ti} = \mathbf{X}_i\boldsymbol{\beta} + \sum_{r=1}^{L} \mathrm{y}_{t-r,i}\gamma_{t-r} + \boldsymbol{Z}_i\boldsymbol{\theta} + e_i$$

where the teacher effect (θ) is a random effect so that it is assumed that

$$\boldsymbol{\theta} \sim N(0, \sigma_\theta^2)$$

and σ_θ^2 is the (fitted) variance of the teacher effects, y_{ti} is the observed score at time *t* for student *i*, $\mathbf{X}_i$ is the *i*th row of the model matrix for the student demographic variables, $\boldsymbol{\beta}$ is a vector of coefficients capturing the effects of the demographic variables included in the model, $y_{t-r,i}$ are the observed lagged scores (in the same tested subject) at time *t–r* ($r \in \{1,2,\dots,L\}$), $\boldsymbol{\gamma}$ is the coefficient vector capturing the effects of lagged scores, $\mathbf{Z}_i$ is a design matrix with one column for each teacher. The entries in the $\mathbf{Z}$ matrix indicate the association between the student test score represented in the row and the teachers represented in the column. The value-added score

[88] To account for the errors in the right hand side variables, we subtracted off the variance due to measurement error from the design matrix, and to account for the measurement error in the left hand side variables, we adjusted the residual term (Doran 2014).

for each teacher (θ) was generated based on the empirical Bayes estimate from the random-effects model.

Covariates Included in the Models for Each District

A set of common covariates were included in the value-added models for all study districts: achievement scores from two prior years (where available, within the same subject), missing data indicators for those prior scores, and fixed effects for the number of relevant courses (minus 1) that a student took for a given subject and grade.[89]

Beyond those common covariates, districts in the study were offered the choice of a selection of non-achievement covariates to include in their value-added model. The "menu" of covariates included the following:

- Special education status (or student disability codes)
- Student differential age (from the expected age for a grade level)
- Free/reduced price meal status (or economically disadvantaged status)
- Prior year attendance/absences
- Student mobility
- Student suspensions
- Class size
- Race/Ethnicity
- Gender
- English Language Learner status

We asked the districts which of these covariates they wanted to include in their value-added model, whether or not they had the data to support the inclusion of the covariates, and at which level(s) they wanted to model the covariate. For example, districts could choose to include special education status as a student-level covariate and/or include the percentage of students with disabilities as a teacher/classroom-level covariate in the value-added model. Districts varied in their selection of covariates, with some districts chose not to include any student demographics in the model.

Calculation of Teacher Overall Value-Added Scores, School Value-Added Scores, and District Value-Added Scores

Because our model generated value-added scores that were teacher-, subject-, grade-, and year-specific, we aggregated the value-added scores for teachers teaching multiple grades and/or subjects to produce an overall value-added score for each teacher for each school year. We also aggregated teacher value-added scores to produce school-level and district-level value-added scores presented in the student growth reports for principals. Below we describe the process of calculating teacher overall value-added scores and school/district value-added scores, which were obtained for each year separately.

[89] We controlled for the number of relevant courses a student took in the same subject and grade because students who took more courses in the same subject and grade were likely to learn more than students who took fewer relevant courses.

To produce an overall value-added score for each teacher for each year, we first standardized the teacher/subject/grade-specific value-added scores for that year within subject, grade, and district based on the standard deviation in the student test scores. We then calculated the variance of the standardized value-added scores using the Taylor series approximation—also called Fieller's Method (Fieller, 1954). Next, we calculated the year-specific overall value-added score for each teacher by averaging across all the subjects and grades the teacher taught in that year, with weights proportional to the inverse variance of the value-added score for a given subject and grade.

The computation of the variance of the overall value-added score for each teacher was complicated by the fact that there could be covariance among the subject/grade/year-specific value-added scores for teachers if a teacher taught the same students in both math and reading in a given year. When this happens, the covariance term would not be zero and was approximated within teacher with

$$cov(\delta_{g,math}, \delta_{g,read}) \approx p_{gs} cov(\hat{r}_{g,math}, \hat{r}_{g,read}),$$

where r is the residual of the fixed portion of the regression $(\mathrm{y}_{ti} - (\mathbf{X}_i\hat{\boldsymbol{\beta}} + \sum_{r=1}^{L} \mathrm{y}_{t-r,i}\hat{\gamma}_{t-r}))$, and $p_g = \frac{n_{j\ common}}{n_{g\ math} \times n_{g\ read}}$ where n is the number of students in reading, math, or common between the two, depending on the subscript. Both the covariance and the value of $p_g$were calculated at the teacher/grade level.

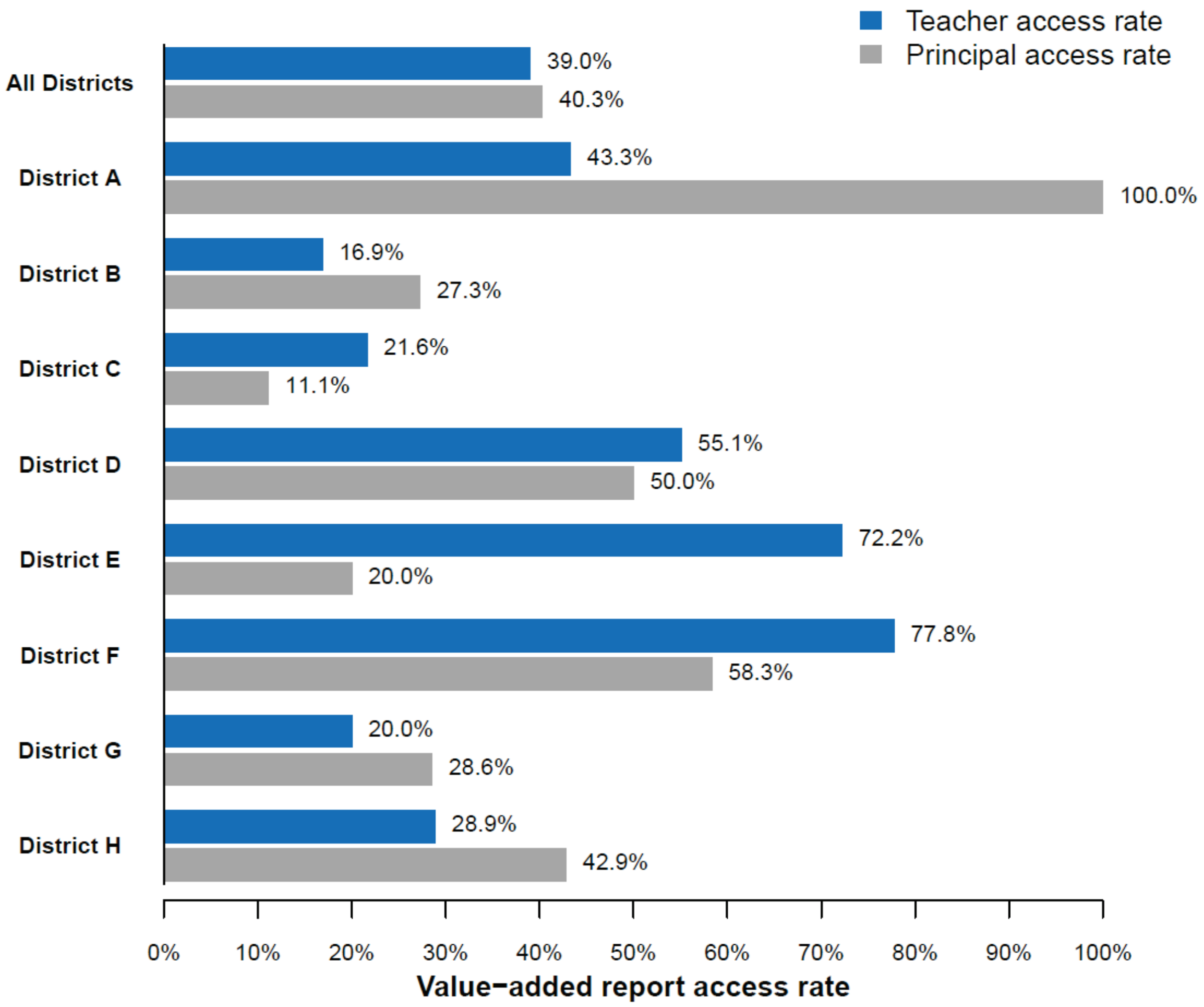
To obtain school value-added scores for a given year, we first calculated a set of subject- and grade-specific value-added scores for each school as information-weighted average of non-standardized teacher value-added scores. We also estimated the variance of these subject- and grade-specific school value-added scores using the covariance terms across teachers from the random-effects regression. We then followed the same steps outlined above for computing teacher overall value-added scores to obtain the school value-added scores aggregated across subjects and grades. District value-added scores were obtained using similar procedures.

The procedures described above calculated the value-added scores at the teacher, school, and district levels for each school year separately. In the student growth reports provided to teachers as part of the study's intervention, a teacher's overall value-added score averaged across the current year and the prior year with information weighting was reported if the teacher had value-added scores from both years; otherwise the teacher's score in the report would be based on value-added data from a single year. The school value-added scores and district value-added scores presented in the student growth reports for principals are also information-weighted two-year averages. The student growth reports provided to principals also include simple unweighted school and district averages of teacher value-added scores, which are intended to allow the principal to compare an individual teacher's performance to the performance of the average teacher in the school or district.

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

Appendix G. Supplemental Findings About the Implementation of the Intervention's Measure of Student Growth

Exhibit G.1. Value-added report access rates for teachers and principals, by district



NOTE: Sample size = 433 teachers and 62 schools.
SOURCE: AIR value-added system.

Exhibit G.2. Descriptive statistics for value-added scores, by teacher characteristics

	Overall value-added score			Reading/ELA value-added score			Mathematics value-added score		
	N	**Mean**	**Standard deviation**	***N***	**Mean**	**Standard deviation**	***N***	**Mean**	**Standard deviation**
All teachers	433	0.01	0.12	326	0.00	0.09	342	0.02	0.18
Grade level									
4 and 5	227	0.00	0.14	208	0.00	0.11	211	0.02	0.20
6–8	206	0.01	0.09	118	0.01	0.06	131	0.01	0.12
Subject taught									
General	266	0.00	0.12	244	0.00	0.10	248	0.01	0.19
Mathematics	89	0.03	0.14	*NA*	*NA*	*NA*	88	0.03	0.14
Reading/ELA	78	0.01	0.05	77	0.01	0.05	*NA*	*NA*	*NA*
Years of experience									
0–3	43	-0.02	0.10	32	-0.03	0.09	36	-0.01	0.18
4–10	148	0.00	0.12	110	0.00	0.11	122	0.01	0.18
11–20	117	0.02	0.12	87	0.01	0.08	88	0.05	0.18
20+	115	0.01	0.11	89	0.01	0.08	90	0.01	0.16

NOTE: *NA* = not applicable.

SOURCE: AIR value-added system.

Exhibit G.3. Distribution of treatment teachers based on their subject area value-added scores being considered measurably above or below the district average

Reading/ELA score	**Mathematics score**		
	Measurably below average	Not measurably different from average	Measurably above average
Measurably below average	7.1%	5.0%	0.0%
Not measurably different from average	17.2%	37.2%	21.3%
Measurably above average	0.8%	3.4%	8.0%

NOTE: The distribution of teachers is based on whether the 80 percent confidence interval for a teacher's value-added score in reading/ELA and mathematics was above or below the district average. Sample size = 239 teachers.

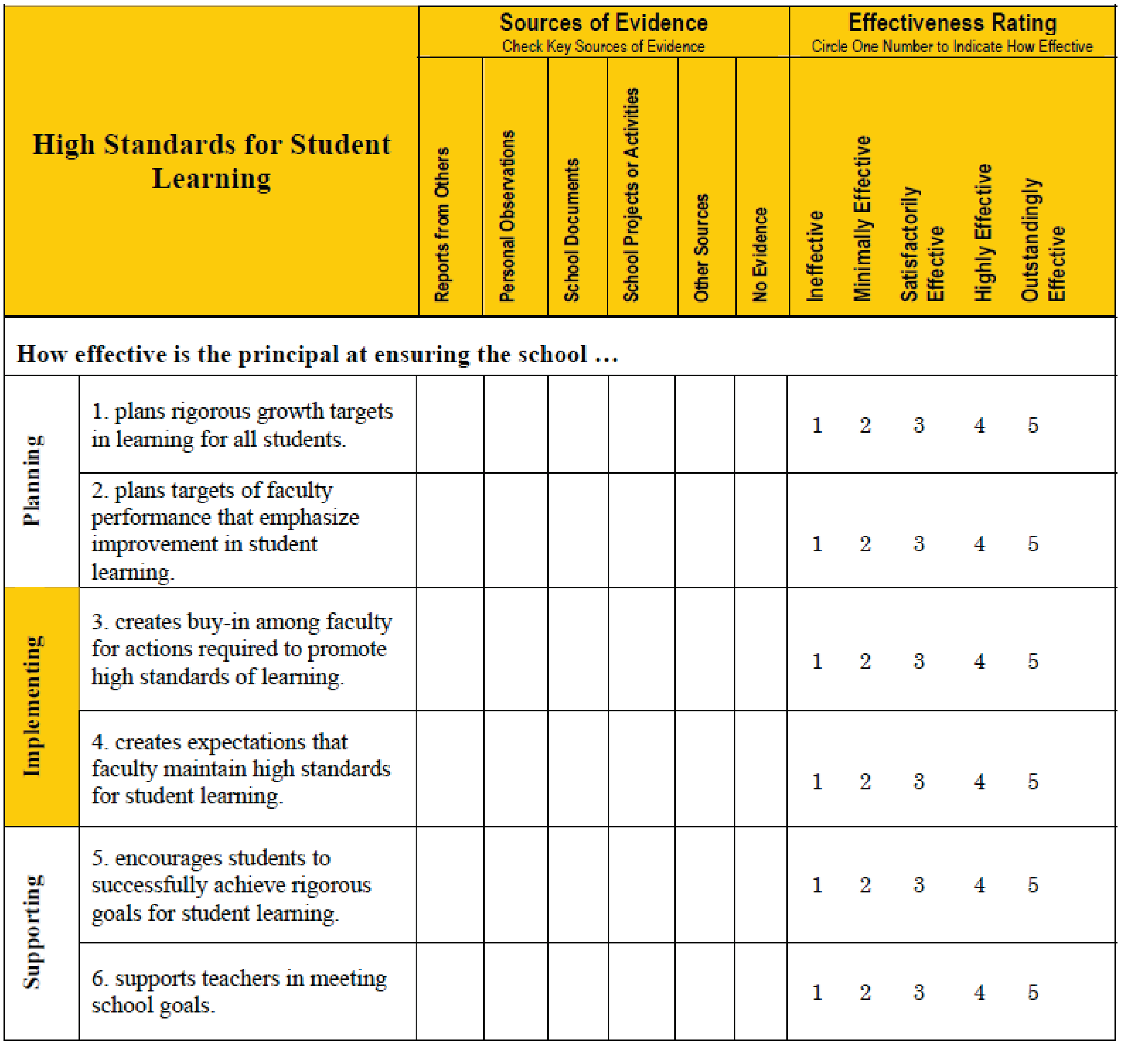
SOURCE: AIR value-added system.

Appendix H. Supplemental Findings About the Intervention's Measure of Principal Leadership

Exhibit H.1. Definitions of VAL-ED core components and key processes

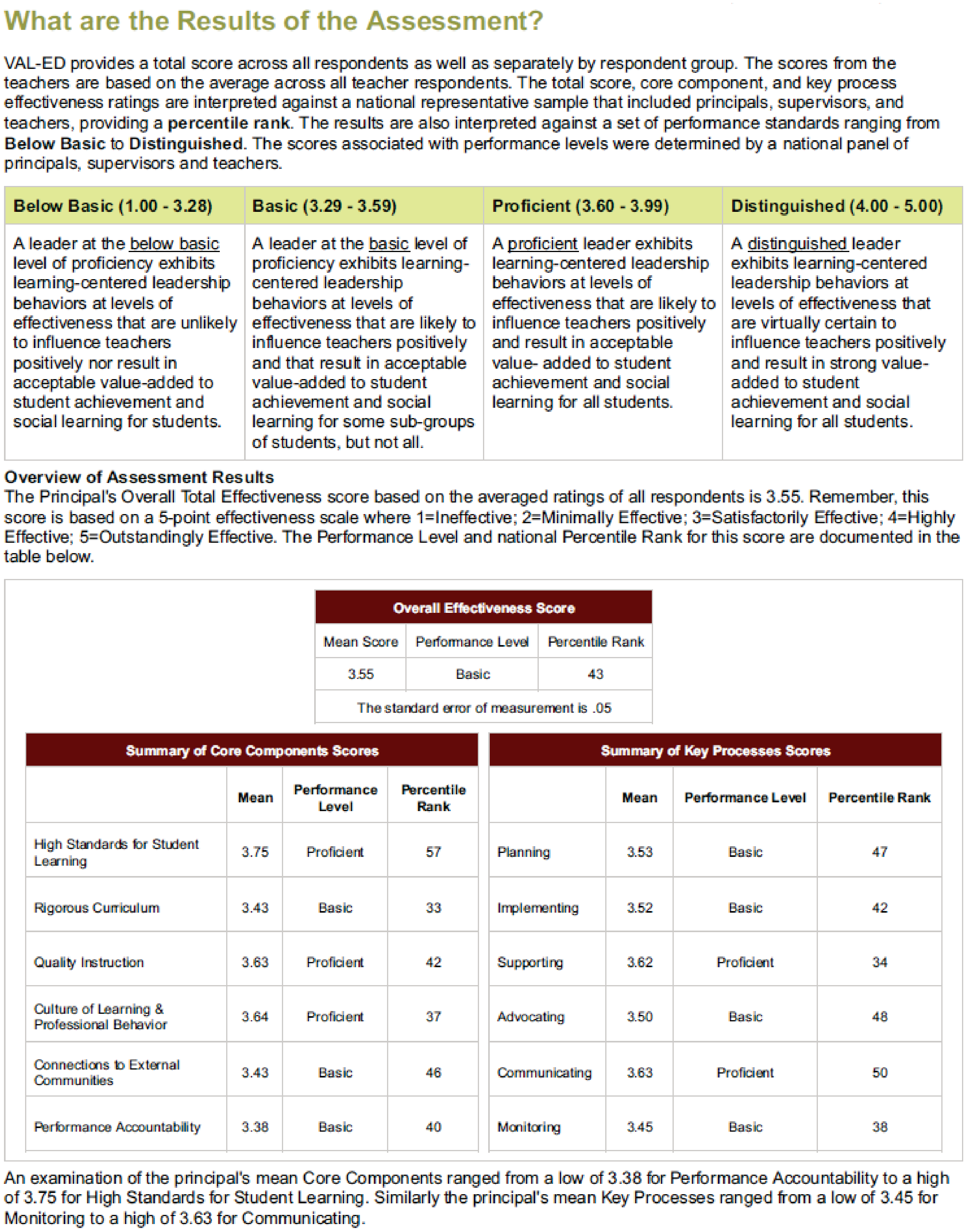
Component or process	Definition
Core components	
High standards for student learning	The school leader ensures there are individual, team, and school goals for rigorous student academic and social learning.
Rigorous curriculum	The school leader ensures ambitious academic content is provided to all students in core academic subjects.
Quality instruction	The school leader ensures effective instructional practices maximize student academic and social learning.
Culture of learning and professional behavior	The school leader ensures there are integrated communities of professional practice in the service of student academic and social learning—that is, a healthy school environment in which student learning is the central focus.
Connections to external communities	The school leader ensures robust connections to the external community.
Systemic performance accountability	The school leader ensures individual and collective responsibility among the leadership, faculty, students, and the community for achieving the rigorous student academic and social learning goals.
Key processes	
Planning	The school leader articulates shared directions and coherent policies, practices, and procedures for realizing high standards of student performance.
Implementing	The school leader engages people, ideas, and resources to put into practice the activities necessary to realize high standards for student performance.
Supporting	The school leader creates enabling conditions; secures and uses the financial, political, technological, and human resources necessary to promote academic and social learning.
Advocating	The school leader promotes the diverse needs of students within and beyond the school.
Communicating	The school leader develops, utilizes, and maintains systems of exchange among members of the school and external communities.
Monitoring	The school leader systematically collects and analyzes data to make judgments that guide decisions and actions.

Exhibit H.2. Sample VAL-ED survey items



High Standards for Student Learning		Sources of Evidence Check Key Sources of Evidence						Effectiveness Rating Circle One Number to Indicate How Effective				
		Reports from Others	Personal Observations	School Documents	School Projects or Activities	Other Sources	No Evidence	Ineffective	Minimally Effective	Satisfactorily Effective	Highly Effective	Outstandingly Effective
How effective is the principal at ensuring the school …												
Planning	1. plans rigorous growth targets in learning for all students.							1	2	3	4	5
	2. plans targets of faculty performance that emphasize improvement in student learning.							1	2	3	4	5
Implementing	3. creates buy-in among faculty for actions required to promote high standards of learning.							1	2	3	4	5
	4. creates expectations that faculty maintain high standards for student learning.							1	2	3	4	5
Supporting	5. encourages students to successfully achieve rigorous goals for student learning.							1	2	3	4	5
	6. supports teachers in meeting school goals.							1	2	3	4	5

Exhibit H.3. Results overview from a sample VAL-ED report



What are the Results of the Assessment?

VAL-ED provides a total score across all respondents as well as separately by respondent group. The scores from the teachers are based on the average across all teacher respondents. The total score, core component, and key process effectiveness ratings are interpreted against a national representative sample that included principals, supervisors, and teachers, providing a **percentile rank**. The results are also interpreted against a set of performance standards ranging from **Below Basic** to **Distinguished**. The scores associated with performance levels were determined by a national panel of principals, supervisors and teachers.

Below Basic (1.00 - 3.28)	Basic (3.29 - 3.59)	Proficient (3.60 - 3.99)	Distinguished (4.00 - 5.00)
A leader at the below basic level of proficiency exhibits learning-centered leadership behaviors at levels of effectiveness that are unlikely to influence teachers positively nor result in acceptable value-added to student achievement and social learning for students.	A leader at the basic level of proficiency exhibits learning-centered leadership behaviors at levels of effectiveness that are likely to influence teachers positively and that result in acceptable value-added to student achievement and social learning for some sub-groups of students, but not all.	A proficient leader exhibits learning-centered leadership behaviors at levels of effectiveness that are likely to influence teachers positively and result in acceptable value- added to student achievement and social learning for all students.	A distinguished leader exhibits learning-centered leadership behaviors at levels of effectiveness that are virtually certain to influence teachers positively and result in strong value-added to student achievement and social learning for all students.

Overview of Assessment Results
The Principal's Overall Total Effectiveness score based on the averaged ratings of all respondents is 3.55. Remember, this score is based on a 5-point effectiveness scale where 1=Ineffective; 2=Minimally Effective; 3=Satisfactorily Effective; 4=Highly Effective; 5=Outstandingly Effective. The Performance Level and national Percentile Rank for this score are documented in the table below.

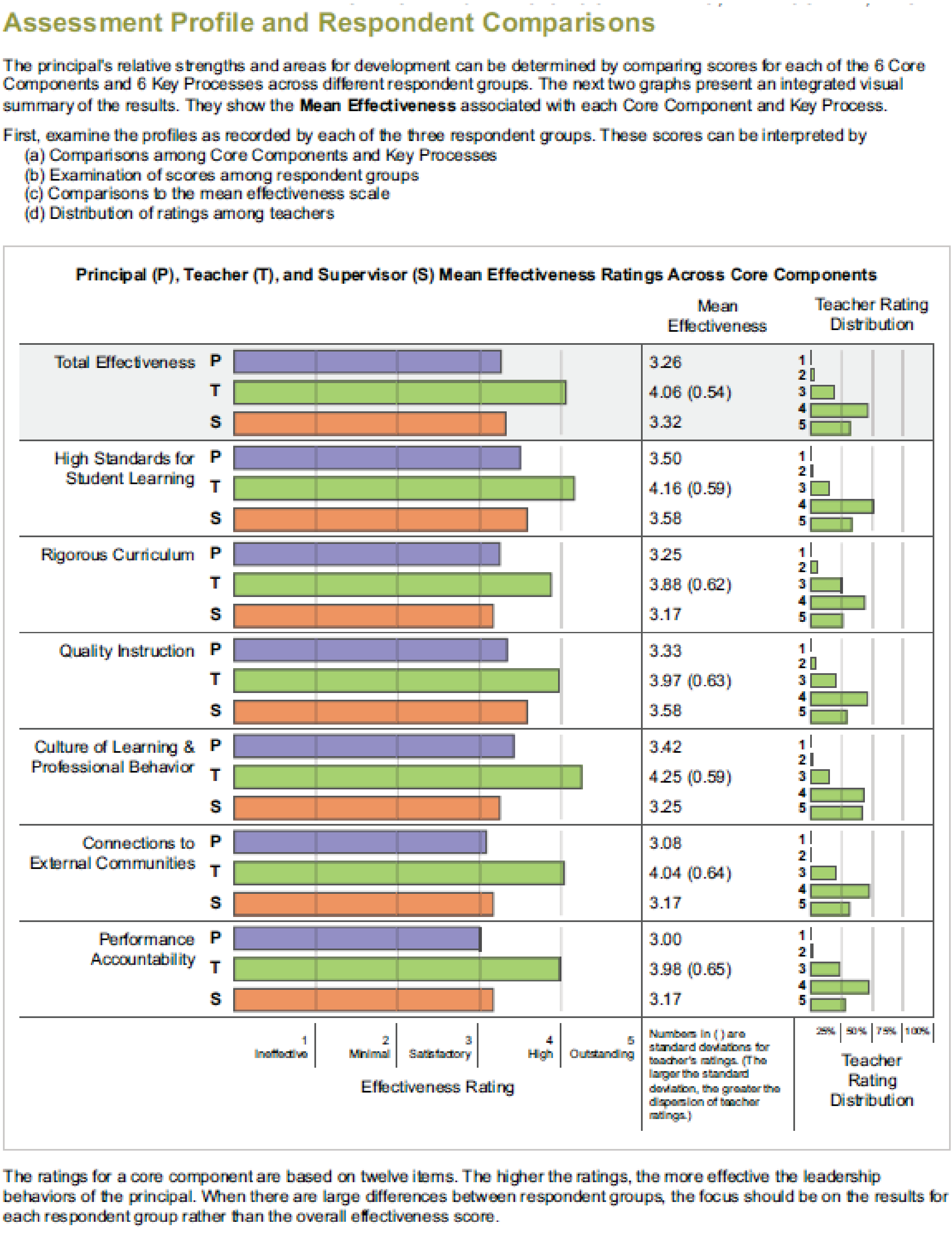
Overall Effectiveness Score		
Mean Score	Performance Level	Percentile Rank
3.55	Basic	43
The standard error of measurement is .05		

Summary of Core Components Scores			
	Mean	Performance Level	Percentile Rank
High Standards for Student Learning	3.75	Proficient	57
Rigorous Curriculum	3.43	Basic	33
Quality Instruction	3.63	Proficient	42
Culture of Learning & Professional Behavior	3.64	Proficient	37
Connections to External Communities	3.43	Basic	46
Performance Accountability	3.38	Basic	40

Summary of Key Processes Scores			
	Mean	Performance Level	Percentile Rank
Planning	3.53	Basic	47
Implementing	3.52	Basic	42
Supporting	3.62	Proficient	34
Advocating	3.50	Basic	48
Communicating	3.63	Proficient	50
Monitoring	3.45	Basic	38

An examination of the principal's mean Core Components ranged from a low of 3.38 for Performance Accountability to a high of 3.75 for High Standards for Student Learning. Similarly the principal's mean Key Processes ranged from a low of 3.45 for Monitoring to a high of 3.63 for Communicating.

Exhibit H.4. Results by respondent group from a sample VAL-ED report



Assessment Profile and Respondent Comparisons

The principal's relative strengths and areas for development can be determined by comparing scores for each of the 6 Core Components and 6 Key Processes across different respondent groups. The next two graphs present an integrated visual summary of the results. They show the **Mean Effectiveness** associated with each Core Component and Key Process.

First, examine the profiles as recorded by each of the three respondent groups. These scores can be interpreted by
(a) Comparisons among Core Components and Key Processes
(b) Examination of scores among respondent groups
(c) Comparisons to the mean effectiveness scale
(d) Distribution of ratings among teachers

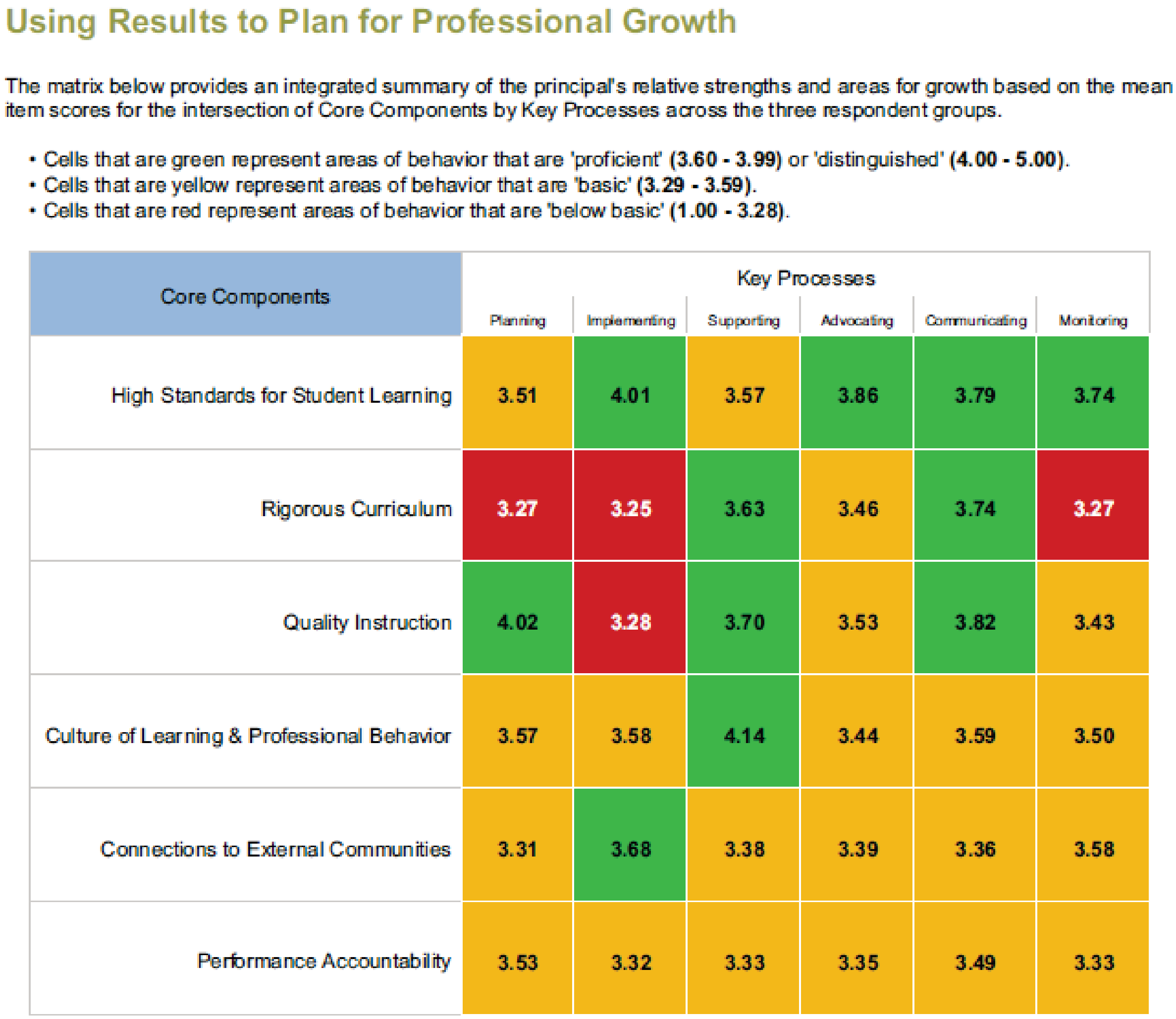
Principal (P), Teacher (T), and Supervisor (S) Mean Effectiveness Ratings Across Core Components

		Mean Effectiveness
Total Effectiveness	P	3.26
	T	4.06 (0.54)
	S	3.32
High Standards for Student Learning	P	3.50
	T	4.16 (0.59)
	S	3.58
Rigorous Curriculum	P	3.25
	T	3.88 (0.62)
	S	3.17
Quality Instruction	P	3.33
	T	3.97 (0.63)
	S	3.58
Culture of Learning & Professional Behavior	P	3.42
	T	4.25 (0.59)
	S	3.25
Connections to External Communities	P	3.08
	T	4.04 (0.64)
	S	3.17
Performance Accountability	P	3.00
	T	3.98 (0.65)
	S	3.17

Numbers in ( ) are standard deviations for teacher's ratings. (The larger the standard deviation, the greater the dispersion of teacher ratings.)

The ratings for a core component are based on twelve items. The higher the ratings, the more effective the leadership behaviors of the principal. When there are large differences between respondent groups, the focus should be on the results for each respondent group rather than the overall effectiveness score.

Exhibit H.5. Summary of component-by-process scores from a sample VAL-ED report



Using Results to Plan for Professional Growth

The matrix below provides an integrated summary of the principal's relative strengths and areas for growth based on the mean item scores for the intersection of Core Components by Key Processes across the three respondent groups.

- Cells that are green represent areas of behavior that are 'proficient' **(3.60 - 3.99)** or 'distinguished' **(4.00 - 5.00)**.
- Cells that are yellow represent areas of behavior that are 'basic' **(3.29 - 3.59)**.
- Cells that are red represent areas of behavior that are 'below basic' **(1.00 - 3.28)**.

Core Components	Key Processes					
	Planning	Implementing	Supporting	Advocating	Communicating	Monitoring
High Standards for Student Learning	3.51	4.01	3.57	3.86	3.79	3.74
Rigorous Curriculum	3.27	3.25	3.63	3.46	3.74	3.27
Quality Instruction	4.02	3.28	3.70	3.53	3.82	3.43
Culture of Learning & Professional Behavior	3.57	3.58	4.14	3.44	3.59	3.50
Connections to External Communities	3.31	3.68	3.38	3.39	3.36	3.58
Performance Accountability	3.53	3.32	3.33	3.35	3.49	3.33

Exhibit H.6. Descriptive statistics for VAL-ED scores, by dimension, fall 2012

	Mean	Standard deviation	Minimum	25th percentile	75th percentile	Maximum
Overall score	3.46	0.32	2.73	3.26	3.67	4.32
Core components						
High standards for student learning	3.54	0.36	2.72	3.29	3.75	4.60
Quality instruction	3.50	0.34	2.68	3.25	3.75	4.34
Culture of learning and professional behavior	3.57	0.33	2.83	3.40	3.77	4.45
Connections to external communities	3.34	0.33	2.32	3.14	3.59	3.99
Performance accountability	3.36	0.37	2.51	3.11	3.63	4.28
Rigorous curriculum	3.45	0.34	2.62	3.22	3.68	4.30
Key processes						
Planning	3.43	0.33	2.72	3.21	3.61	4.29
Implementing	3.48	0.32	2.78	3.28	3.67	4.38
Supporting	3.53	0.32	2.80	3.27	3.70	4.52
Advocating	3.45	0.31	2.64	3.27	3.64	4.23
Communicating	3.46	0.34	2.79	3.22	3.71	4.29
Monitoring	3.44	0.38	2.54	3.21	3.67	4.21

NOTE: Sample size = 63 principals.

SOURCE: Fall 2012 VAL-ED Surveys.

Exhibit H.7. Descriptive statistics for VAL-ED scores, by dimension, spring 2013

	Mean	Standard deviation	Minimum	25th percentile	75th percentile	Maximum
Overall score	3.61	0.35	2.67	3.41	3.86	4.46
Core components						
High standards for student learning	3.67	0.39	2.56	3.35	3.92	4.64
Quality instruction	3.71	0.37	2.67	3.46	3.97	4.62
Culture of learning and professional behavior	3.71	0.37	2.56	3.51	3.97	4.47
Connections to external communities	3.43	0.33	2.44	3.29	3.63	4.24
Performance accountability	3.57	0.43	2.53	3.34	3.85	4.63
Rigorous curriculum	3.58	0.36	2.55	3.37	3.77	4.57
Key processes						
Planning	3.58	0.35	2.75	3.32	3.82	4.43
Implementing	3.61	0.35	2.60	3.41	3.83	4.58
Supporting	3.72	0.37	2.60	3.53	3.96	4.64
Advocating	3.57	0.36	2.55	3.34	3.78	4.40
Communicating	3.60	0.36	2.61	3.38	3.87	4.40
Monitoring	3.58	0.38	2.57	3.35	3.82	4.51

NOTE: Sample size = 63 principals.

SOURCE: Spring 2013 VAL-ED Surveys.

Exhibit H.8. Correlations among the VAL-ED dimension scores, fall 2012

Domain	Dimension of leadership	C1	C2	C3	C4	C5	C6	P1	P2	P3	P4	P5	P6
Core components	(C1) High standards for student learning	1.00											
	(C2) Quality instruction	.90	1.00										
	(C3) Culture of learning and professional behavior	.81	.84	1.00									
	(C4) Connections to external communities	.73	.72	.83	1.00								
	(C5) Performance accountability	.87	.90	.85	.76	1.00							
	(C6) Rigorous curriculum	.92	.90	.84	.76	.89	1.00						
Key processes	(P1) Planning	.92	.93	.89	.82	.93	.91	1.00					
	(P2) Implementing	.92	.92	.90	.81	.92	.91	.95	1.00				
	(P3) Supporting	.90	.94	.88	.76	.91	.91	.94	.94	1.00			
	(P4) Advocating	.84	.86	.92	.86	.88	.88	.90	.89	.85	1.00		
	(P5) Communicating	.89	.87	.89	.85	.88	.91	.89	.90	.86	.88	1.00	
	(P6) Monitoring	.91	.90	.87	.81	.94	.92	.91	.90	.87	.90	.93	1.00

NOTE: Sample size = 63 principals. Shaded cells represent correlations among core components or key processes.

SOURCE: Fall 2012 VAL-ED Surveys.

Exhibit H.9. Correlations among the VAL-ED dimension scores, spring 2013

Domain	Dimension of leadership	C1	C2	C3	C4	C5	C6	P1	P2	P3	P4	P5	P6
Core components	(C1) High standards for student learning	1.00											
	(C2) Quality instruction	.94	1.00										
	(C3) Culture of learning and professional behavior	.83	.91	1.00									
	(C4) Connections to external communities	.68	.73	.84	1.00								
	(C5) Performance accountability	.91	.95	.89	.73	1.00							
	(C6) Rigorous curriculum	.94	.95	.86	.68	.93	1.00						
Key processes	(P1) Planning	.94	.96	.91	.74	.94	.95	1.00					
	(P2) Implementing	.94	.95	.93	.81	.94	.93	.95	1.00				
	(P3) Supporting	.93	.95	.91	.78	.94	.93	.93	.96	1.00			
	(P4) Advocating	.90	.94	.95	.85	.93	.92	.93	.94	.93	1.00		
	(P5) Communicating	.93	.95	.93	.81	.96	.93	.94	.94	.94	.95	1.00	
	(P6) Monitoring	.91	.95	.91	.77	.95	.93	.93	.93	.91	.94	.95	1.00

NOTE: Sample size = 63 principals. Shaded cells represent correlations among core components or key processes.

SOURCE: Spring 2013 VAL-ED Surveys.

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

Appendix I. Supplemental Findings About Educators' Experiences

Exhibit I.1a. Percentage of teachers who reported receiving ratings on their performance in CLASS districts, by treatment status

CLASS teachers	Treatment group mean	Control group mean	Estimated difference	Standard error	Effect size	*p* value
Overall	78.4	37.4	41.0*	4.8	0.84	.000
Nonprobationary teachers	77.7	32.0	45.6*	5.4	0.97	.000
Probationary teachers	83.3	61.6	21.7*	10.2	0.43	.033

NOTE: Sample size = 63 schools (31 treatment and 32 control) and 629 grades 4–8 teachers (302 treatment and 326 control). The analyses were based on a two-level regression (teachers within schools) controlling for random assignment blocks. Statistically significant difference (*p* < .05, two-tailed) between the treatment and control groups is indicated by an asterisk (*).

SOURCE: Spring 2013 Teacher Survey.

Exhibit I.1b. Percentage of teachers who reported receiving ratings on their performance in FFT districts, by treatment status

FFT teachers	Treatment group mean	Control group mean	Estimated difference	Standard error	Effect size	*p* value
Overall	88.4	39.8	48.6*	4.9	0.98	.000
Nonprobationary teachers	92.3	33.4	58.9*	5.8	1.26	.000
Probationary teachers	87.6	78.5	9.2	11.0	0.21	.403

NOTE: Sample size = 64 schools (32 treatment and 32 control) and 443 grades 4–8 teachers (218 treatment and 225 control). The analyses were based on a two-level regression (teachers within schools) controlling for random assignment blocks. Statistically significant difference (*p* < .05, two-tailed) between the treatment and control groups is indicated by an asterisk (*).

SOURCE: Spring 2013 Teacher Survey.

Exhibit I.1c. Percentage of K–3 teachers who reported receiving ratings on their performance, by treatment status

K–3 teachers	Treatment group mean	Control group mean	Estimated difference	Standard error	Effect size	*p* value
Overall	77.7	40.1	37.6*	3.2	0.77	.000
Nonprobationary teachers	80.0	35.5	44.5*	3.3	0.95	.000
Probationary teachers	67.1	59.8	7.3	7.6	0.15	.340

NOTE: Sample size = 100 schools (50 treatment and 50 control) and 1,072 grades K–3 teachers (523 treatment and 549 control). The analyses were based on a two-level regression (teachers within schools) controlling for random assignment blocks. Statistically significant difference (*p* < .05, two-tailed) between the treatment and control groups is indicated by an asterisk (*).

SOURCE: Spring 2013 Teacher Survey.

Exhibit I.2a. Number of feedback instances and duration of feedback that an average teacher in CLASS districts reported receiving, by treatment status

Feedback	Treatment group median	Control group median	Estimated difference	*p* value
Number of instances with any type of feedback	4.0	3.0	1.0*	.000
Number of feedback sessions with ratings and written narrative	3.0	1.0	2.0*	.000
Total length of oral feedback	60.0	6.5	53.5*	.000

NOTE: Sample size = 63 schools (31 treatment and 32 control) and 629 grades 4–8 teachers (305 treatment and 324 control). The analyses were based on an aligned rank sum test with randomization inference about median difference between treatment and control groups. Statistically significant difference (*p* < .05, two-tailed) between the treatment and control groups is indicated by an asterisk (*).

SOURCE: Spring 2013 Teacher Survey.

Exhibit I.2b. Number of feedback instances and duration of feedback that an average teacher in FFT districts reported receiving, by treatment status

Feedback	Treatment group median	Control group median	Estimated difference	*p* value
Number of instances with any type of feedback	4.0	3.3	0.7*	.000
Number of feedback sessions with ratings and written narrative	3.0	0.2	2.8*	.000
Total length of oral feedback	95.0	19.4	75.6*	.000

NOTE: Sample size = 64 schools (32 treatment and 32 control) and 443 grades 4–8 teachers (218 treatment and 225 control). The analyses were based on an aligned rank sum test with randomization inference about median difference between treatment and control groups. Statistically significant difference (*p* < .05, two-tailed) between the treatment and control groups is indicated by an asterisk (*).

SOURCE: Spring 2013 Teacher Survey.

Exhibit I.2c. Number of feedback instances and duration of feedback that an average K–3 teacher reported receiving, by treatment status

Feedback	Treatment group median	Control group median	Estimated difference	*p* value
Number of instances with any type of feedback	2.0	2.0	0.0	.934
Number of feedback sessions with ratings and written narrative	1.0	0.1	0.9*	.000
Total length of oral feedback	45.0	17.8	27.2*	.000

NOTE: Sample size = 100 schools (50 treatment and 50 control) and 1,072 grades K–3 teachers (523 treatment and 549 control). The analyses were based on an aligned rank sum test with randomization inference about median difference between treatment and control groups. Statistically significant difference (*p* < .05, two-tailed) between the treatment and control groups is indicated by an asterisk (*).

SOURCE: Spring 2013 Teacher Survey.

Exhibit I.3a. Percentage of teachers in CLASS districts who reported discussing areas of classroom practice related to CLASS/FFT with someone who provided them with feedback during the school year, by treatment status

Area of practice	Treatment group mean	Control group mean	Estimated difference	Standard error	Effect size	*p* value
Behavior management	50.7	51.8	-1.1	4.6	-0.02	.819
Classroom organization	45.9	39.9	6.0	4.9	0.12	.227
Emotional support for students	54.6	40.1	14.5*	4.3	0.29	.001
Instructional dialogue	70.7	50.5	20.2*	4.3	0.40	.000
Student engagement	66.1	50.0	16.1*	5.0	0.32	.001

NOTE: Sample size = 63 schools (31 treatment and 32 control) and 544 or 545 grades 4–8 teachers (268–270 treatment and 274–276 control). The analyses were based on a two-level regression (teachers within schools) controlling for random assignment blocks. Statistically significant difference (*p* < .05, two-tailed) between the treatment and control groups is indicated by an asterisk (*).

SOURCE: Spring 2013 Teacher Survey.

Exhibit I.3b. Percentage of teachers in FFT districts who reported discussing areas of classroom practice related to CLASS/FFT with someone who provided them with feedback during the school year, by treatment status

Area of practice	Treatment group mean	Control group mean	Estimated difference	Standard error	Effect size	*p* value
Behavior management	62.0	51.1	10.9	6.3	0.22	.082
Classroom organization	58.8	39.2	19.6*	5.3	0.40	.000
Emotional support for students	46.4	38.4	8.0	5.8	0.16	.168
Instructional dialogue	73.2	58.1	15.1*	5.3	0.31	.005
Student engagement	80.9	55.7	25.1*	5.3	0.51	.000

NOTE: Sample size = 64 schools (32 treatment and 32 control) and 403–405 grades 4–8 teachers (192–194 treatment and 209–212 control). The analyses were based on a two-level regression (teachers within schools) controlling for random assignment blocks. Statistically significant difference (*p* < .05, two-tailed) between the treatment and control groups is indicated by an asterisk (*).

SOURCE: Spring 2013 Teacher Survey.

Exhibit I.3c. Percentage of K–3 teachers who reported discussing areas of classroom practice related to CLASS/FFT with someone who provided them with feedback during the school year, by treatment status

Area of practice	Treatment group mean	Control group mean	Estimated difference	Standard error	Effect size	*p* value
Behavior management	60.1	60.6	-0.5	4.0	-0.01	.891
Classroom organization	53.8	44.5	9.3*	3.9	0.19	.017
Emotional support for students	50.0	41.9	8.1*	4.1	0.17	.047
Instructional dialogue	72.0	53.4	18.6*	3.8	0.37	.000
Student engagement	69.5	55.6	13.9*	3.3	0.28	.000

NOTE: Sample size = 100 schools (50 treatment and 50 control) and 947–950 grades K–3 teachers (460–463 treatment and 485–488 control). The analyses were based on a two-level regression (teachers within schools) controlling for random assignment blocks. Statistically significant difference (*p* < .05, two-tailed) between the treatment and control groups is indicated by an asterisk (*).

SOURCE: Spring 2013 Teacher Survey.

Exhibit I.4a. Percentage of teachers in CLASS districts who reported receiving specific types of student achievement information, by treatment status

Type of information	Treatment group mean	Control group mean	Estimated difference	Standard error	Effect size	*p* value
Value-added scores for me based upon the students that I taught	37.7	29.4	8.4	4.4	0.18	.059
Data on individual students that I taught	63.3	85.4	-22.1*	3.6	-0.62	.000
Average data for classes of students that I taught	49.3	64.3	-15.0*	4.2	-0.31	.000
I did not receive any student achievement information based on standardized test results	15.9	4.9	11.0*	2.8	0.49	.000

NOTE: Sample size = 63 schools (31 treatment and 32 control) and 628 grades 4–8 teachers (302 treatment and 326 control). The analyses were based on a two-level regression (teachers within schools) controlling for random assignment blocks. Statistically significant difference (*p* < .05, two-tailed) between the treatment and control groups is indicated by an asterisk (*).

SOURCE: Spring 2013 Teacher Survey.

Exhibit I.4b. Percentage of teachers in FFT districts who reported receiving specific types of student achievement information, by treatment status

Type of information	Treatment group mean	Control group mean	Estimated difference	Standard error	Effect size	*p* value
Value -added scores for me based upon the students that I taught	51.5	19.5	31.9*	4.5	0.79	.000
Data on individual students that I taught	64.1	82.6	-18.5*	4.2	-0.49	.000
Average data for classes of students that I taught	53.1	60.7	-7.6	4.6	-0.16	.101
I did not receive any student achievement information based on standardized test results	15.0	8.1	6.9	3.6	0.24	.056

NOTE: Sample size = 64 schools (32 treatment and 32 control) and 445 grades 4–8 teachers (217 treatment and 228 control). The analyses were based on a two-level regression (teachers within schools) controlling for random assignment blocks. Statistically significant difference (*p* < .05, two-tailed) between the treatment and control groups is indicated by an asterisk (*).

SOURCE: Spring 2013 Teacher Survey.

Exhibit I.4c. Percentage of K–3 teachers who reported receiving specific types of student achievement information, by treatment status

Type of information	Treatment group mean	Control group mean	Estimated difference	Standard error	Effect size	*p* value
Value-added scores for me based upon the students that I taught	16.5	19.3	-2.8	2.5	-0.06	.269
Data on individual students that I taught	60.0	73.4	-13.4*	2.9	-0.37	.000
Average data for classes of students that I taught	43.7	54.0	-10.3*	3.3	-0.21	.002
I did not receive any student achievement information based on standardized test results	31.3	16.8	14.5*	2.9	0.58	.000

NOTE: Sample size = 100 schools (50 treatment and 50 control) and 1,073 grades K–3 teachers (519 treatment and 554 control). The analyses were based on a two-level regression (teachers within schools) controlling for random assignment blocks. Statistically significant difference (*p* < .05, two-tailed) between the treatment and control groups is indicated by an asterisk (*).

SOURCE: Spring 2013 Teacher Survey.

Exhibit I.5a. Percentage of teachers in CLASS districts who agreed or strongly agreed with statements about the performance feedback they received, by treatment status

Statements	Treatment group mean	Control group mean	Estimated difference	Standard error	Effect size	*p* value
Feedback was a fair assessment of my performance	93.5	92.0	1.5	2.2	0.05	.493
Feedback included specific ideas about how I could improve my performance	91.4	82.5	9.0*	3.3	0.24	.007

NOTE: Sample size = 63 schools (31 treatment and 32 control) and 583–587 grades 4–8 teachers (301–303 treatment and 282–284 control). The analyses were based on a two-level regression (teachers within schools) controlling for random assignment blocks. Statistically significant difference (*p* < .05, two-tailed) between the treatment and control groups is indicated by an asterisk (*).

SOURCE: Spring 2013 Teacher Survey.

Exhibit I.5b. Percentage of teachers in FFT districts who agreed or strongly agreed with statements about the performance feedback they received, by treatment status

Statements	Treatment group mean	Control group mean	Estimated difference	Standard error	Effect size	*p* value
Feedback was a fair assessment of my performance	89.8	90.5	-0.7	3.3	-0.02	.831
Feedback included specific ideas about how I could improve my performance	82.4	75.8	6.6	4.7	0.15	.158

NOTE: Sample size = 64 schools (32 treatment and 32 control) and 421 grades 4–8 teachers (218 treatment and 203 control). The analyses were based on a two-level regression (teachers within schools) controlling for random assignment blocks. Statistically significant difference (*p* < .05, two-tailed) between the treatment and control groups is indicated by an asterisk (*).

SOURCE: Spring 2013 Teacher Survey.

Exhibit I.5c. Percentage of K–3 teachers who agreed or strongly agreed with statements about the performance feedback they received, by treatment status

Statements	Treatment group mean	Control group mean	Estimated difference	Standard error	Effect size	*p* value
Feedback was a fair assessment of my performance	93.2	93.4	-0.3	1.6	-0.01	.864
Feedback included specific ideas about how I could improve my performance	88.2	83.0	5.2*	2.6	0.13	.048

NOTE: Sample size = 100 schools (50 treatment and 50 control) and 1,004–1,008 grades K–3 teachers (519–521 treatment and 485–487 control). The analyses were based on a two-level regression (teachers within schools) controlling for random assignment blocks. Statistically significant difference (*p* < .05, two-tailed) between the treatment and control groups is indicated by an asterisk (*).

SOURCE: Spring 2013 Teacher Survey.

Exhibit I.6a. Percentage of teachers in CLASS districts who agreed or strongly agreed with statements about the rating system used for the majority of the ratings they received, by treatment status

Survey item	Treatment group mean	Control group mean	Estimated difference	Standard error	Effect size	*p* value
The rating system does a good job distinguishing effective from ineffective teaching.	82.0	82.6	-0.5	5.0	-0.01	.919
I have a clear idea of what the rating system views as "good instruction."	91.5	90.0	1.5	3.7	0.05	.696
The way my teaching is being rated accurately reflects the quality of my teaching.	78.4	82.9	-4.5	5.2	-0.12	.384
The rating system is fair to all teachers, regardless of their personal characteristics or those of the students they teach	72.4	81.2	-8.8	6.1	-0.23	.151

NOTE: Sample size = 355–357 teachers (235–239 treatment and 118–121 control). The analyses were based on a two-level regression (teachers within schools) controlling for random assignment blocks. None of the differences between the treatment and the control groups were statistically significant at the .05 level (two-tailed).

SOURCE: Spring 2013 Teacher Survey.

Exhibit I.6b. Percentage of teachers in FFT districts who agreed or strongly agreed with statements about the rating system used for the majority of the ratings they received, by treatment status

Survey item	Treatment group mean	Control group mean	Estimated difference	Standard error	Effect size	*p* value
The rating system does a good job distinguishing effective from ineffective teaching.	73.9	79.6	-5.7	6.0	-0.14	.340
I have a clear idea of what the rating system views as "good instruction."	81.1	81.1	0.0	6.1	0.00	.996
The way my teaching is being rated accurately reflects the quality of my teaching.	75.2	81.5	-6.3	6.2	-0.16	.309
The rating system is fair to all teachers, regardless of their personal characteristics or those of the students they teach	62.5	79.2	-16.7*	6.8	-0.38	.014

NOTE: Sample size = 276–282 teachers (184–189 treatment and 91–93 control). The analyses were based on a two-level regression (teachers within schools) controlling for random assignment blocks. None of the differences between the treatment and the control groups were statistically significant at the .05 level (two-tailed).

SOURCE: Spring 2013 Teacher Survey.

Exhibit I.6c. Percentage of K–3 teachers who agreed or strongly agreed with statements about the rating system used for the majority of the ratings they received, by treatment status

Survey item	Treatment group mean	Control group mean	Estimated difference	Standard error	Effect size	*p* value
The rating system does a good job distinguishing effective from ineffective teaching.	84.1	78.3	5.8	3.4	0.15	.090
I have a clear idea of what the rating system views as “good instruction.”	90.4	85.3	5.1	2.9	0.16	.082
The way my teaching is being rated accurately reflects the quality of my teaching.	82.5	83.9	-1.5	3.6	-0.04	.686
The rating system is fair to all teachers, regardless of their personal characteristics or those of the students they teach	78.0	78.8	-0.8	4.2	-0.02	.850

NOTE: Sample size = 631–639 teachers (419–428 treatment and 211 or 212 control). The analyses were based on a two-level regression (teachers within schools) controlling for random assignment blocks. None of the differences between the treatment and the control groups were statistically significant at the .05 level (two-tailed).

SOURCE: Spring 2013 Teacher Survey.

Exhibit I.7a. Percentage of teachers in CLASS districts who agreed or strongly agreed with statements about the fairness of the student achievement information they received, by treatment status

Statements	Treatment group mean	Control group mean	Estimated difference	Standard error	Effect size	*p* value
The information is fair to all teachers, regardless of the personal characteristics of the students they teach	47.6	32.3	15.3*	4.1	0.32	.000
The information is fair to all teachers, regardless of the prior achievement of the students they teach	48.8	30.7	18.1*	4.6	0.39	.000
The information is a fair assessment of my performance	59.1	47.2	11.9*	4.2	0.24	.005
The information is easy to understand	84.9	89.0	-4.2	2.8	-0.13	.135

NOTE: Sample size = 63 schools (31 treatment and 32 control) and 561–565 grades 4–8 teachers (254–256 treatment and 305–310 control). The analyses were based on a two-level regression (teachers within schools) controlling for random assignment blocks. Statistically significant difference (*p* < .05, two-tailed) between the treatment and control groups is indicated by an asterisk (*).

SOURCE: Spring 2013 Teacher Survey.

Exhibit I.7b. Percentage of teachers in FFT districts who agreed or strongly agreed with statements about the fairness of the student achievement information they received, by treatment status

Statements	Treatment group mean	Control group mean	Estimated difference	Standard error	Effect size	*p* value
The information is fair to all teachers, regardless of the personal characteristics of the students they teach	32.4	26.5	5.8	5.0	0.13	.241
The information is fair to all teachers, regardless of the prior achievement of the students they teach	35.4	24.7	10.6*	4.9	0.24	.030
The information is a fair assessment of my performance	40.2	39.9	0.2	6.5	0.00	.970
The information is easy to understand	72.1	88.8	-16.7*	4.5	-0.55	.000

NOTE: Sample size = 64 schools (32 treatment and 32 control) and 387–389 grades 4–8 teachers (181–184 treatment and 205–207 control). The analyses were based on a two-level regression (teachers within schools) controlling for random assignment blocks. Statistically significant difference (*p* < .05, two-tailed) between the treatment and control groups is indicated by an asterisk (*).

SOURCE: Spring 2013 Teacher Survey.

Exhibit I.7c. Percentage of K–3 teachers who agreed or strongly agreed with statements about the fairness of the student achievement information they received, by treatment status

Statements	Treatment group mean	Control group mean	Estimated difference	Standard error	Effect size	*p* value
The information is fair to all teachers, regardless of the personal characteristics of the students they teach	47.9	29.0	18.9*	3.5	0.41	.000
The information is fair to all teachers, regardless of the prior achievement of the students they teach	48.7	30.4	18.3*	3.9	0.40	.000
The information is a fair assessment of my performance	57.5	45.0	12.5*	4.2	0.25	.003
The information is easy to understand	93.3	91.2	2.1	2.4	0.07	.378

NOTE: Sample size = 100 schools (50 treatment and 50 control) and 949–954 grades K–3 teachers (437–439 treatment and 512–515 control). The analyses were based on a two-level regression (teachers within schools) controlling for random assignment blocks. Statistically significant difference (*p* < .05, two-tailed) between the treatment and control groups is indicated by an asterisk (*).

SOURCE: Spring 2013 Teacher Survey.

Exhibit I.8. Percentage of principals who reported discussing with their supervisors areas unrelated to VAL-ED, by treatment status

Improvement area	Treatment group mean	Control group mean	Estimated difference	Standard error	Effect size	*p* value
Making personnel/human resources decisions	54.5	53.9	0.6	8.7	0.01	0.945
Managing non-personnel administrative issues (e.g., budgeting, facilities maintenance)	32.7	37.6	-4.9	7.9	-0.10	0.539
Student behavior/discipline (e.g., drug/crime prevention; social development)	30.9	41.0	-10.1	8.5	-0.20	0.239

NOTE: Sample size = 123 principals (61 treatment and 62 control). The analyses were based on a principal-level regression controlling for random assignment blocks. None of the differences between the treatment and the control groups were statistically significant at the .05 level (two-tailed).

SOURCE: Spring 2013 Principal Survey.

Exhibit I.9. Difference between probationary and nonprobationary teachers in the treatment-control difference in teachers' performance evaluation experience

Treatment effect on:	Chapter 5 exhibit with overall finding	Estimated difference by probationary status	Standard error	*p* value
Percentage of teachers reporting receiving ratings on classroom practice	Exhibit 5.1	-34.7*	7.0	0.000
Number of instances of any type of feedback	Exhibit 5.2	-0.3	0.2	0.121
Number of feedback sessions with ratings and written narrative	Exhibit 5.2	-0.4*	0.2	0.049
Total length of oral feedback (minute)	Exhibit. 5.2	-20.7*	7.0	0.003
Percentage of teachers reporting discussing classroom practice areas related to CLASS/FFT and areas not related, with someone providing them with feedback during the school year	Exhibit 5.3			
Behavior management		7.7	8.7	0.372
Classroom organization		-9.7	8.7	0.264
Emotional support		-4.5	8.8	0.611
Instructional dialogue		-16.7*	8.4	0.046
Student engagement		-9.2	8.3	0.268
Lesson planning		-0.2	8.8	0.986
Data use		13.1	8.6	0.127
Content-specific teaching techniques		-8.8	8.8	0.316
Content knowledge		8.2	8.8	0.351
Percentage of teachers reporting receiving specific type of student achievement information	Exhibit 5.4			
Value-added scores based on students that I taught		-0.6	7.8	0.935
Data on individual students that I taught		10.6	7.1	0.135
Average data for classes of students that I taught		7.1	8.2	0.383
Percentage of teachers agreeing or strongly agreeing with statements about the performance feedback they received	Exhibit 5.5			
Feedback was a fair assessment of my performance.		3.5	4.7	0.451
Feedback was easy to understand.		2.8	3.5	0.428
Feedback included specific ideas about how I could improve my performance.		4.5	6.1	0.464
The feedback made me more reflective about my teaching.		1.6	5.6	0.780
In the long run, students will benefit from the feedback I received.		4.1	5.7	0.474

Exhibit I.9. Difference between probationary and nonprobationary teachers in the treatment-control difference in teachers' performance evaluation experience (continued)

Treatment effect on:	Chapter 5 exhibit with overall finding	Estimated difference by probationary status	Standard error	p value
Percentage of teachers agreeing or strongly agreeing with statements about rating systems	Exhibit 5.6			
The rating system does a good job distinguishing effective from ineffective teaching.		-7.1	9.0	0.433
I have a clear idea of what the rating system views as "good instruction."		1.8	7.3	0.801
The way my teaching is being rated accurately reflects the quality of my teaching.		5.4	9.2	0.555
The rating system is fair to all teachers, regardless of their personal characteristics or those of the students they teach.		-18.3	9.8	0.062
Percentage of teachers agreeing or strongly agreeing with statements about the fairness of student achievement information they received	Exhibit 5.7			
The information is easy to understand.		-5.8	6.4	0.372
The information is fair to all teachers, regardless of the personal characteristics of the students they teach.		-11.3	8.6	0.187
The information is fair to all teachers, regardless of the prior achievement of the students they teach.		-11.4	8.7	0.188
The information is a fair assessment of my performance.		-17.2	9.1	0.060

NOTE: See relevant exhibits in chapter 5 for sample size information. All analyses were based on two-level linear regression models (teachers within schools) controlling for random assignment blocks. Estimated difference by probationary status represents the difference between the treatment effect on teachers' performance evaluation experience among probationary teachers and the treatment effect among nonprobationary teachers. A positive estimate indicates a larger treatment effect among probationary teachers relative to nonprobationary teachers. Statistically significant difference (*p* < .05, two-tailed) by probationary status is indicated by an asterisk (*).

SOURCE: Spring 2013 Teacher Survey.

Exhibit I.10a. Teacher background characteristics of teachers who received feedback based on a classroom observation, by study group

Characteristic	Treatment group	Control group	Estimated difference	*p* value
Probationary teacher (percentage)	18.6	24.6	-5.9	.076
Years of experience				
Mean number of years	13.6	13.7	-0.2	.815
Three years or fewer (percentage)	13.0	18.0	-5.0	.058
Three to 10 years (percentage)	35.7	29.3	6.4	.058
Ten to 20 years (percentage)	26.7	27.4	-0.8	.804
More than 20 years (percentage)	24.7	25.1	-0.4	.893
Master's degree or higher (percentage)	45.9	43.6	2.4	.427
Number of teachers	**521**	**489**		

NOTE: Sample size for master's degree of higher = 520 treatment and 548 control). The analyses are based on a two-level linear regression model controlling for random assignment blocks. The treatment group means are unadjusted means, and the control group means were computed by subtracting the estimated group differences from the unadjusted treatment group means. *p* Values are based on *t* tests. Two-tailed statistical significance at the *p* < .05 level is indicated by an asterisk (*).

SOURCE: Spring 2013 Teacher Survey

Exhibit I.10b. Teacher background characteristics of teachers who received a rating based on a classroom observation, by study group

Characteristic	Treatment group	Control group	Estimated difference	*p* value
Probationary teacher (percentage)	18.7	39.1	-20.4	.000
Years of experience				
Mean number of years	13.9	12.1	1.8	.058
Three years or fewer (percentage)	12.8	26.3	-13.5	.001
Three to 10 years (percentage)	34.5	31.9	2.6	.597
Ten to 20 years (percentage)	26.5	20.3	6.2	.155
More than 20 years (percentage)	26.2	21.9	4.3	.279
Master's degree or higher (percentage)	46.3	40.5	5.8	.128
Number of teachers	**428**	**216**		

NOTE: Sample size for master's degree of higher = 427 treatment and 216 control. The analyses are based on a two-level linear regression model controlling for random assignment blocks. The treatment group means are unadjusted means, and the control group means were computed by subtracting the estimated group differences from the unadjusted treatment group means. *p* Values are based on *t* tests. Two-tailed statistical significance at the *p* < .05 level is indicated by an asterisk (*).

SOURCE: Spring 2013 Teacher Survey

Exhibit I.10c. Teacher background characteristics of teachers who viewed student achievement information, by study group

Characteristic	Treatment group	Control group	Estimated difference	*p* value
Probationary teacher (percentage)	18.0	20.0	-2.0	.544
Years of experience				
Mean number of years	13.7	14.2	-0.5	.491
Three years or fewer (percentage)	12.5	14.7	-2.2	.406
Three to 10 years (percentage)	35.7	30.0	5.7	.096
Ten to 20 years (percentage)	26.5	28.9	-2.4	.442
More than 20 years (percentage)	25.3	26.1	-0.8	.809
Master's degree or higher (percentage)	47.8	44.1	3.6	.221
Number of teachers	**444**	**521**		

NOTE: Sample size for master's degree of higher = 443 treatment and 520 control. The analyses are based on a two-level linear regression model controlling for random assignment blocks. The treatment group means are unadjusted means, and the control group means were computed by subtracting the estimated group differences from the unadjusted treatment group means. *p* Values are based on *t* tests. Two-tailed statistical significance at the *p* < .05 level is indicated by an asterisk (*).

SOURCE: Spring 2013 Teacher Survey

Exhibit I.10d. Percentage of nonprobationary teachers who agreed or strongly agreed with statements about the rating system used for the majority of the ratings they received, by treatment status

Survey item	Treatment group mean	Control group mean	Estimated difference	Standard error	Effect size	*p* value
The rating system does a good job distinguishing effective from ineffective teaching.	76.9	75.9	1.0	5.2	0.03	0.850
I have a clear idea of what the rating system views as "good instruction."	85.3	85.9	-0.7	4.5	-0.02	0.884
The way my teaching is being rated accurately reflects the quality of my teaching.	76.6	82.4	-5.8	5.1	-0.16	0.259
The rating system is fair to all teachers, regardless of their personal characteristics or those of the students they teach	67.2	71.8	-4.6	6.0	-0.11	0.443

NOTE: Sample size = 482–490 teachers (345–351 treatment and 137–140 control). The analyses were based on a two-level regression (teachers within schools) controlling for random assignment blocks. None of the differences between the treatment and the control groups were statistically significant at the .05 level (two-tailed).

SOURCE: Spring 2013 Teacher Survey.

Exhibit I.10e. Principal background characteristics of principals who received feedback from their supervisor, by study group

Characteristic	Treatment group	Control group	Estimated difference	*p* value
Years of experience as a principal				
Mean number of years	7.9	9.4	-1.5	.402
Three years or fewer (percentage)	26.0	15.3	10.7	.383
Three to 10 years (percentage)	44.9	54.2	-9.3	.601
Ten to 20 years (percentage)	25.7	23.2	2.5	.872
More than 20 years (percentage)	3.5	7.3	-3.9	.648
Years of experience as a teacher				
Mean number of years	12.5	10.4	2.1	.168
Three years or fewer (percentage)	0.0	4.0	-4.0	.117
Three to 10 years (percentage)	45.0	50.6	-5.6	.667
Ten to 20 years (percentage)	45.0	42.1	2.9	.854
More than 20 years (percentage)	10.0	3.3	6.6	.441
Number of teachers	**53**	**36**		

NOTE: The analyses were based on a principal-level regression controlling for random assignment blocks.. *p* Values are based on *t* tests. Two-tailed statistical significance at the $p < .05$ level is indicated by an asterisk (*).

SOURCE: Spring 2013 Principal Survey

Appendix J. Sample Reports

Sample CLASS Observation Report

Teachstone Building connections. Enhancing learning. CLASS

CLASS™ Classroom Report

Teacher:	Teacher B
School:	School P
Grade Level:	4
Subject:	Mathematics
Observation:	**3**
Date:	**02/22/2013**

This report summarizes CLASS observation results from your classroom. The CLASS observation measures effective teacher-student interactions. Please refer to your Dimensions Guide for more information.

This report provides the following information:

- **Section I:** Summary of the current observation.
- **Section II:** Detailed information and observation notes from the current observation.
- **Section III:** Summary of all observations to date.

Section I: Observation 3 Summary

Date	Emotional Support	Classroom Organization	Instructional Support	Student Engagement	Overall Score*
02/22/2013	5.16 ▂▄▆█	6.33 ▂▄▆█	3.9 ▂▄▆	5.5 ▂▄▆█	4.95 ▂▄▆

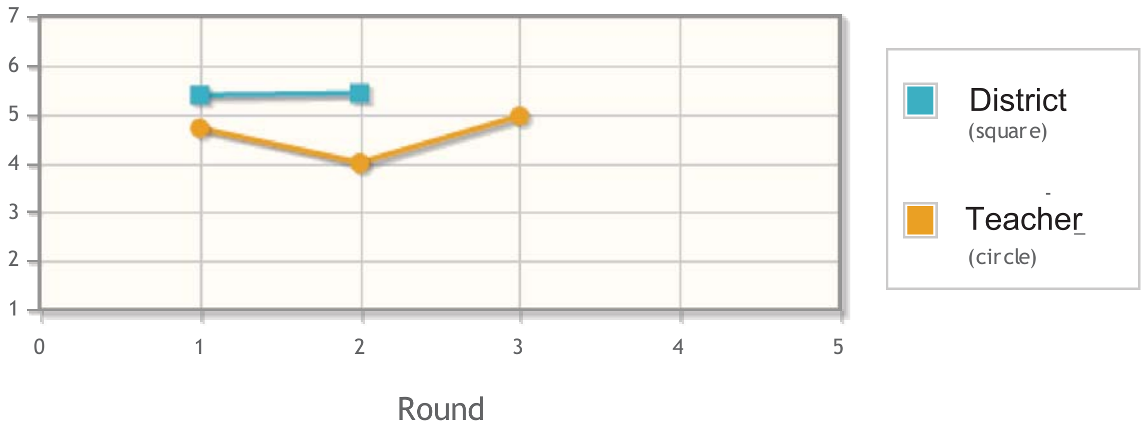
Key: ▂ Ineffective ▂▄ Developing Effectiveness ▂▄▆ Effective ▂▄▆█ Highly Effective

*The Overall Score is calculated by averaging all dimensions. Note: The mapping of CLASS scores onto effectiveness categories varies by domain.

Context of the Observation:

The Round 3 observation began when the class had just returned from an activity in the Computer Lab. The students put their notebooks away and were given an opportunity to enjoy a quick snack and some social conversation as they had flexibility to move about the room in a relaxed format before their math lesson began. When Teacher B gave the signal, the students gathered their math materials and sat on the floor in the front of the room to correct and discuss their homework assignment. Moving on, the students reviewed the Identity Property of Addition and Multiplication. The class discussed how to use the Identity Property to simplify an equation. Examples were given. Discussions took place involving the inverse operations of multiplication and addition and variables. Independent practice time was given while students had an opportunity to share their results and discussion how they figured out the value of each expression. The observation ended as the students prepared for their daily recess/lunch time .

Overall CLASS Score



*The district average includes only classrooms that received a CLASS observation.

Category	Point Range
Highly Effective	5.00 - 7.00
Effective	3.50 - 4.99
Developing Effectiveness	2.50 - 3.49
Ineffective	1.00 - 2.49

CLASS Advisor Summary

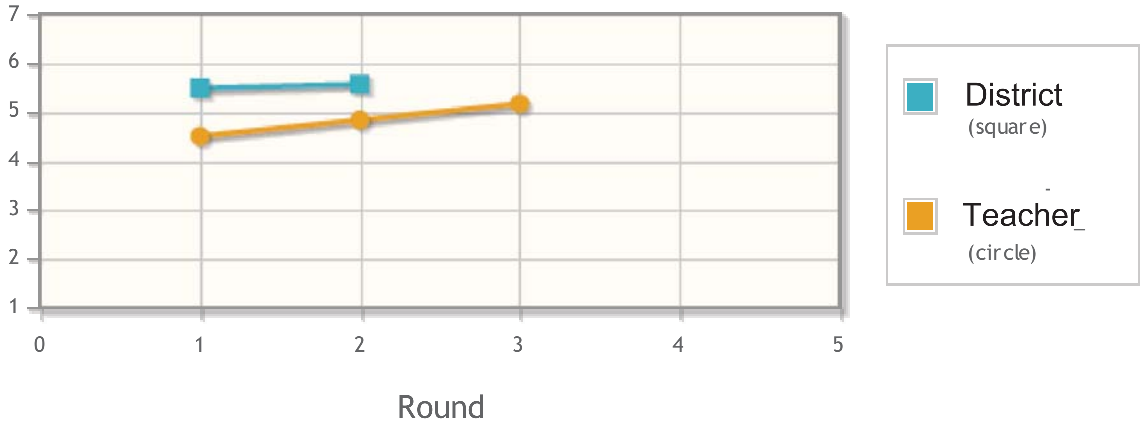
Your overall score was in the **Effective** range. Your areas of strength were indicated in the Emotional Support and Classroom Organization domains as well as Student Engagement.You scored in the highly effective range in these domains however there is always room for continued learning. You demonstrated very effective interactions in Positive Climate and Teacher Sensitivity. Less effective interactions were displayed in Regard for Student Perspectives. Classroom Organization was strong in all three dimensions of Behavior Management, Productivity, and absences of Negative Climate. Although strong and effective in the Instructional Support domain, there were some less effective interactions in Quality of Feedback, and Instructional Learning Formats.

Conference Summary

The Round 3 conference began with a brief discussion of Teacher B's overall CLASS score. We looked at the CLASS Advisor Summary in all three domains focusing on strengths and areas to continue to grow and develop. We discussed the dimension of Regard for Student Perspectives and focused our attention on the indicators of Support for Autonomy/ Leadership and Flexibility /Student Focus...allowing students to lead a lesson and being flexible in ones plans to follow students' lead and instruct around their interest. We viewed video # 3 Giving Students Chances to Lead in a Science Lesson. We paid close attention to the Focus Text for the Clip as well. Moving on we discussed the Instructional Support domain. We covered each dimension and discussed Quality of Feedback indicators. We viewed video # 6 Giving Specific Feedback to Students to Their Presentation. As the conference was coming to its end, we also viewed Behavior Management video # 2 Paying Attention to the Positive Before a Lesson noticing how to be proactive in behavior management to remind and reinforce ones expectations. We also discussed the value in reviewing the Upper Elementary Dimension Guide not only to refresh ones knowledge of the indicators but also to read the tips to promote and develop each particular dimension. The following videos are suggested to view independently. Regard for Student Perspectives Video # 8 Incorporating Students' Points of View into a Summary of the Activity. Quality of Feedback video # Engaging in Feedback Loops in a Math Activity. Behavior Management video #6 Clearly Establishing Expectations Before an Activity Begins.

Section II: Observation 3 Details

Emotional Support Domain



*The district average includes only classrooms that received a CLASS observation.

Category	Point Range
Highly Effective	5.00 - 7.00
Effective	4.00 - 4.99
Developing Effectiveness	3.00 - 3.99
Ineffective	1.00 - 2.99

Class Advisor Summary

Your lesson was marked by **Highly Effective** Emotional Support. Your areas of strength included Positive Climate and Teacher Sensitivity. There were many indications of teacher respect and positive affect among you and your students. You offered one-on-one instructional support and responded to students needs. Although it fell in the effective range, Regard for Student Perspectives is an area of focus. In the CLASS video library, under RSP, please consider viewing video # 3 Giving Students Chances to Lead in a Science Lesson. Notice how the teacher promotes student lead presentations and allows students to ask questions to their peers.The teacher places emphasis on students' ideas and encourages student responsibility and autonomy.

Video recommendations for this domain:

- http://class.teachstone.com/video_library/video_ue/vid_detail.php?id=167

Emotional Support Dimensions

Positive Climate 6.0

Highly Effective. There was very strong evidence of effective Positive Climate in your classroom.

During the observation, the following effective examples were noted:

- Teacher B demonstrated respect by calling his students by name, speaking in a calm voice, and using respectful language which included "Please"and "Thank you" responses.

- As the class was correcting their independent practice examples, there were some displays of matched positive affect of excitement to go to the smartboard to complete a math problem, displays of smiles, and some giggles when selecting students to share their work.

During the observation, the following less effective examples were noted:

- There were indications of blurting and students talking over each other while individuals had the floor to participate and share their ideas.

Teacher Sensitivity 5.5

Highly Effective. There was very strong evidence of effective Teacher Sensitivity in your classroom.

During the observation, the following effective examples were noted:

- There were frequent indications that the fourth grade students responded to Teacher B's questions and participated in the lesson.
- As the students worked independently, Teacher B offered one-on-one support to students who were struggling with their task to use the identity property to simplify an expression.

During the observation, the following less effective examples were noted:

- When correcting the previous night's homework assignment, there was a missed opportunity to acknowledge and assist a student who called out , "I don't understand the clock stuff", during the time allotted to correct homework.

Regard for Student Perspectives 4.0

Effective. There was strong evidence of effective Regard for Student Perspectives in your classroom.

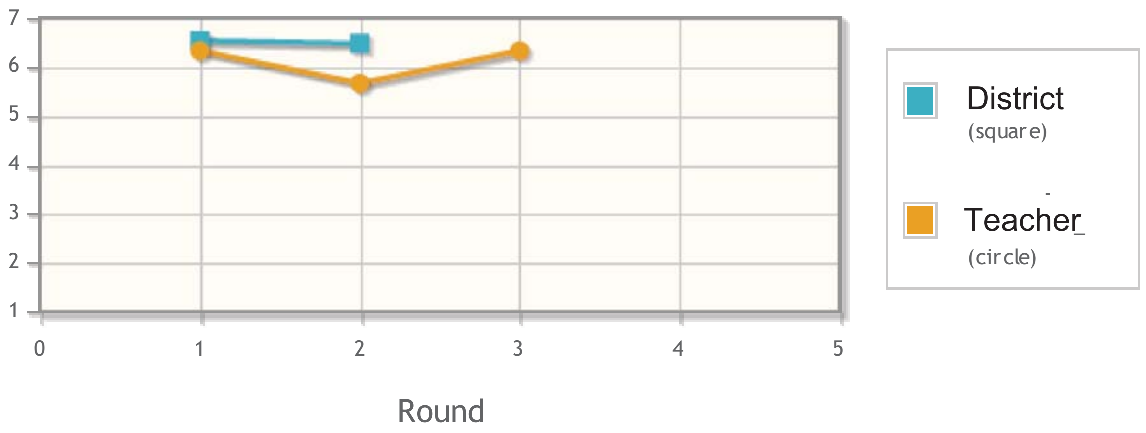
During the observation, the following effective examples were noted:

- During the math review of properties and algebraic notation, the students were given responsibilities to complete practice problems in a relaxed setting.
- Although students worked independently on their practice examples, there was some evidence of meaningful peer exchanges as students discussed math concepts and findings.

During the observation, the following less effective examples were noted:

- The lesson was designed and managed by Teacher B in such a way that the students' opportunity for academic choice or leadership responsibilities was lacking.

Classroom Organization Domain



*The district average includes only classrooms that received a CLASS observation.

Category	Point Range
Highly Effective	6.00 - 7.00
Effective	5.50 - 5.99
Developing Effectiveness	5.00 - 5.49
Ineffective	1.00 - 4.99

Class Advisor Summary

Your lesson was marked by **Highly Effective** Classroom Organization. You were strong in all three dimension of Classroom Organization. There was no evidence of negative climate in your observation. Your areas of strength were Productivity and Behavior Management. The fourth graders were provided with tasks and you were prepared for the lesson. The students followed directions and were responsive to redirection when necessary. There is always room for growth. In the Behavior Management video library, please consider watching video # 2 Paying Attention to the Positive Before a Lesson. Notice how the teacher encourages desirable behavior before starting the lesson to prevent misbehavior. Rather than reacting to misbehavior, she is paying attention to desirable behavior.

Video recommendations for this domain:

- http://class.teachstone.com/video_library/video_ue/vid_detail.php?id=159

Classroom Organization Dimensions

Behavior Management 6.0

Highly Effective. There was very strong evidence of effective Behavior Management in your classroom.

During the observation, the following effective examples were noted:

- Throughout the math activity, the students followed directions and knew what to do while completing their math assignment.
- Teacher B used effective redirection strategies to keep students on task and compliant with the volume in the classroom before it escalated or became an issue in this relaxed work environment.

During the observation, the following less effective examples were noted:

- Clear expectations for sharing math answers/results were not stated at the start of the activity so Teacher B was reactive to their calling out of responses when he said, "Hold on, Hold on, Please stop talking!" " No one can hear with all this calling out."

Productivity 6.0

Highly Effective. There was very strong evidence of effective Productivity in your classroom.

During the observation, the following effective examples were noted:

- The fourth graders demonstrated that they knew what was expected of them through established routines when engaged in whole group and individual formats.
- Teacher B was prepared, knew the subject matter,and had all materials ready and accessible for the students and himself.

During the observation, the following less effective examples were noted:

- Tasks were provided throughout the math time. As the students completed each activity section of the assignment on properties and algebraic notation, Teacher B did not offer a choice when finished before others. Students were told to "wait quietly" while other peers finished up to join in.

Negative Climate 1.0

* For Negative Climate, lower scores indicate more effective interactions. Note that Negative Climate scores are reversed when calculating domain scores.

Highly Effective. There was little or no evidence of Negative Climate in your classroom.

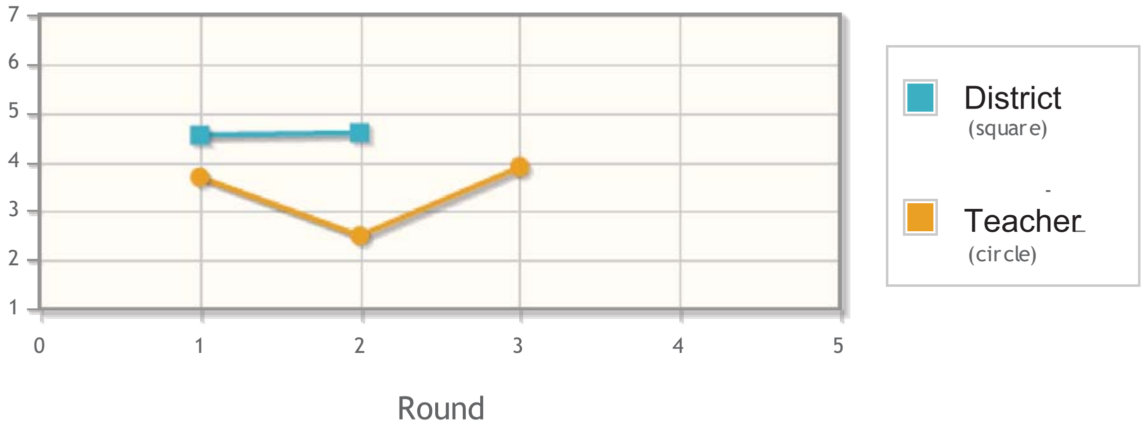
During the observation, the following effective examples were noted:

- There was no evidence of negative affect or disrespect.
- There was no evidence of punitive control.

During the observation, the following less effective examples were noted:

- None were observed during this observation.

Instructional Support Domain



*The district average includes only classrooms that received a CLASS observation.

Category	Point Range
Highly Effective	4.00 - 7.00
Effective	3.00 - 3.99
Developing Effectiveness	2.00 - 2.99
Ineffective	1.00 - 1.99

Class Advisor Summary

Your lesson was marked by **Effective** Instructional Support. Your areas of strength and evidence of Instructional Support were in the dimensions of Content Understanding, where you provided supervised and independent practice time, and Analysis and Inquiry, where you demonstrated metacognition and provided opportunity for higher order thinking skills.An area to focus your attention for continued growth would be in the Quality of Feedback dimension. Please consider viewing Quality of Feedback video # 6 Giving Specific Feedback to Students on Their Presentation. Although the video is very short, notice how the teacher goes beyond simply saying "Good Job". The teacher provides brief but specific feedback about what the students did well.

Video recommendations for this domain:

- http://class.teachstone.com/video_library/video_ue/vid_detail.php?id=72

Instructional Support Dimensions

Instructional Learning Formats 4.0

Highly Effective. There was very strong evidence of effective Instructional Learning Formats in your classroom.

During the observation, the following effective examples were noted:

- Learning objective were discussed. Math information and concepts were presented in a clear format. Students were shown numerous examples of simplifying expressions. Time was spent discussing the importance of the equal sign.
- Teacher B demonstrated active facilitation by promoting participation and showing interest in the students' work.

During the observation, the following less effective examples were noted:

- The students had few opportunities to interact with a variety of materials other than paper pencil tasks in order to complete the assignment. There was a very brief moment to interact with the Smartboard for a select few students who wrote their math answer next to the equations but did not offer any explanation regarding it.

Content Understanding 4.0

Highly Effective. There was very strong evidence of effective Content Understanding in your classroom.

During the observation, the following effective examples were noted:

- Teacher B quickly but clearly demonstrated and communicated the concepts and procedures to be used in solving equations using the identity property to simplify each expression given. He also explained the proper steps on how to evaluate the equation by substituting the value of each letter first and the simplifying the expression.
- The students were provided with supervised and independent practice time of procedures and skills as they completed a worksheet from the curriculum.

During the observation, the following less effective examples were noted:

- Although students applied their background knowledge of math facts, there were no attempts to encourage a deeper understanding of the concepts through real world connections.

Analysis and Inquiry 4.0

Highly Effective. There was very strong evidence of effective Analysis and Inquiry in your classroom.

During the observation, the following effective examples were noted:

- While Teacher B was explaining the identity property to simplify an expression, he modeled his thinking about thinking (metacognition) as he walked through the procedure with the students."The problem is n+5n= 6n. Ok, first I need to find the value of "n". Then I notice that 6n means 6 x any number. If "n" is 1 then 1+ 5 x1= 6 x 1. When I complete the equation I see that 1 + 5 =6 Now I see that 6=6 and I am right."
- With his guidance and support, Teacher B made attempts to ask his students higher order thinking skills by asking students to explain a variety of questions. Explain the identity property of addition and multiplication. Explain what makes an equation. He also asked students to explain the inverse operation of multiplication and how it will help to solve one particular math problem.

During the observation, the following less effective examples were noted:

- Although Teacher B was carrying the cognitive load of the discussions, the examples, and the procedures, he did make attempts to challenge the fourth graders to think about the math concepts.

Quality of Feedback 3.5

Effective. There was strong evidence of effective Quality of Feedback in your classroom.

During the observation, the following effective examples were noted:

- When working one-on-one with students, Teacher B offered hints and gave assistance to students in order to complete the assignment with guided success.
- In large group and during individual support, Teacher B, although brief, used follow up questions to increase student awareness and understanding to math procedures especially when discussing elapsed time examples.

During the observation, the following less effective examples were noted:

- There was occasional evidence of recognition of effort but it was at a perfunctory level and did not increase involvement or effect persistence in the lesson. "Good" "Good job" "OK" "Nice job" .

Instructional Dialogue 4.0

Highly Effective. There was very strong evidence of effective Instructional Dialogue in your classroom.

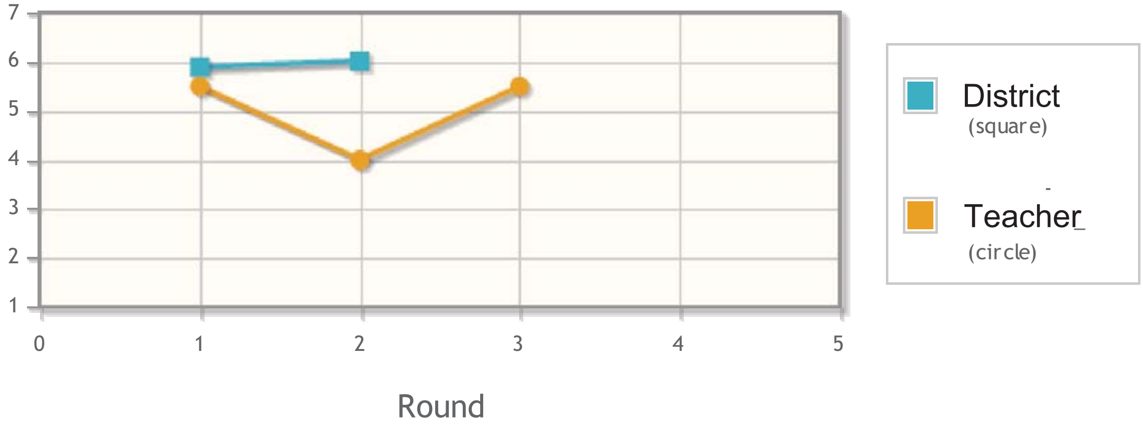
During the observation, the following effective examples were noted:

- There were opportunities for content focused discussions between Teacher B and his fourth grade students. Evaluate, inverse, operation, and simplify were defined and connected to the tasks and conversations often.
- Although not stated or encouraged directly, Teacher B allowed some peer to peer dialogues to support content understanding while students were working on their individual practice time.

During the observation, the following less effective examples were noted:

- The class was mostly dominated by teacher talk but there were instances in which the fourth graders took on more initiative to participate in the discussions and the correcting of the assigned tasks. There were some students who, although alert and aware of the objectives and tasks, never took a verbal role in the activity.

Student Engagement Domain



*The district average includes only classrooms that received a CLASS observation.

Category	Point Range
Highly Effective	5.50 - 7.00
Effective	4.50 - 5.49
Developing Effectiveness	3.50 - 4.49
Ineffective	1.00 - 3.49

Class Advisor Summary

Your lesson was marked by **Highly Effective** Student Engagement. This dimension was an area of strength. The students were engaged, responded to questions and participated during the lesson. Continue to look for the passive students or distracted students in the classroom and engage them in the discussions and activities as well. In the CLASS video library under Student Engagement consider viewing video # 4 Active Engagement in a Discussion about Germs. Notice how the teacher enthusiastically engages the students in a discussion about places one would encounter germs. Notice how the students actively volunteer to share ideas.

Video recommendations for this domain:

- http://class.teachstone.com/video_library/video_ue/vid_detail.php?id=134

Student Engagement Dimensions

Student Engagement 5.5

Highly Effective. There was very strong evidence of effective Student Engagement in your classroom.

During the observation, the following effective examples were noted:

- The fourth graders responded to Teacher B's questions in both whole group and small group formats as he involved students in the homework discussion and independent assignment.
- Some students volunteered to share their math findings while others sat passively listening and observing rather than actively engaging in the activity.

During the observation, the following less effective examples were noted:

- There was some evidence of students disengaged and not participating in the homework discussion or correcting because they did not return their homework assignment. There were no adjustments made to engage them in the activity except to have them follow along without it.

Section III: Summary of Observations to Date

This table summarizes your CLASS observations from all completed observations.

Observation	Date	Emotional Support	Classroom Organization	Instructional Support	Student Engagement	Overall Score*
#1	11/12/2012	4.5 ▂▄▆	6.33 ▂▄▆█	3.7 ▂▄▆	5.5 ▂▄▆█	4.7 ▂▄▆
#2	12/19/2012	4.83 ▂▄▆	5.66 ▂▄▆	2.5 ▂▄	4.0 ▂▄	4.0 ▂▄▆
#3	02/22/2013	5.16 ▂▄▆█	6.33 ▂▄▆█	3.9 ▂▄▆	5.5 ▂▄▆█	4.95 ▂▄▆
Cumulative Average		4.83 ▂▄▆	6.11 ▂▄▆█	3.36 ▂▄▆	5.0 ▂▄▆	4.55 ▂▄▆

Key: ▂ Ineffective ▂▄ Developing Effectiveness ▂▄▆ Effective ▂▄▆█ Highly Effective

*The Overall Score is calculated by averaging all dimensions. Note: The mapping of CLASS scores onto effectiveness categories varies by domain.

Sample FFT Observation Report

Teacher: [redacted]		Title: 6th grade
Scheduled on:	Feb 27, 2013 - 4:46 AM	
Observation date:	Feb 27, 2013 - 4:45 AM	
Submitted by: Jeske, Jim	Mar 03, 2013 - 3:03 PM	
Date Confirmed:	Mar 05, 2013 - 10:12 AM	
Focus:		
Additional instructions:		

Scores and Evidence

2a: Creating an environment of respect and rapport

[redacted] **Score: 3**

Evidence

S- talk in small groups...listening to the student intently.
4:47 am

T- called students by name to share (R and R)
5:04 am

T- "It's pretty nasty isn't it?" -responding to a student cause and effect (smiling)
5:05 am

O- teacher and students smiling during the conversation (R and R)
5:06 am

O- quiet, calm atmosphere...only hear the student reading in small group with Mrs. Overbeck.
5:08 am

T- "What do you think" S- responded T- "way to go, I was thinking the same thing?"
5:12 am

T- "Alright, thank you." Students left the table.
5:15 am

S- made a big circle with notebooks and pencils ready to go. (procedures) T- "I'm impressed" responding to the making of the circle.
5:17 am

O- discussion was respectful...student to student conversations were good supporting whether they were for zoos or not for zoos.
5:20 am

T- "Levi?" have you gone to our zoo?" Are you as guilty as I am for throwing corn at the animals?" S- responded with a smile. (this was in response to a student response to a question)
5:30 am

Critical Attributes

Proficient - Talk between teacher and students and among students is uniformly respectful.
Proficient - Teacher responds to disrespectful behavior among students.
Proficient - Teacher makes superficial connections with individual students.

Summary

You have created a positive, productive classroom environment.

2b: Establishing a culture for learning

[redacted] **Score: 3**

Evidence

S- talk in small groups...listening to the student intently.
4:47 am

O- student sharing his thoughts about the zoo issue...other students listened and jotted down questions to ask at the end.
4:52 am

O- quiet, calm atmosphere...only hear the student reading in small group with Mrs. Overbeck.
5:08 am

T- "I will demonstrate how this will work." (model)
5:18 am

T- "It's not an argument, it's a discussion." (set the table for the group conversation)
5:19 am

Critical Attributes

Proficient - The teacher communicates the importance of learning, and that with hard work all students can be successful in it.
Proficient - The teacher demonstrates a high regard for student abilities.
Proficient - Teacher conveys an expectation of high levels of student effort.
Proficient - Students expend good effort to complete work of high quality.

Summary

You have created clear learning expectations. Your students respectfully share their evidence of learning.

2c: Managing classroom procedures

[redacted] **Score: 3**

Evidence

T- used Actiboard as a timer for the class.
4:46 am

T- "3, 2, 1...next person go."
4:49 am

O- procedures were in place for groups...picked a card. (procedures)
4:52 am

T- "I need all eyes and ears" T- "We will discuss whole group later, right now we are going to do our Daily."
4:55 am

S- made choices in less than 30 seconds. (procedures)
4:56 am

S- checked out to the bathroom using the classroom system.
4:57 am

O- student came back into the room from the bathroom with no disturbances. (procedures)
4:59 am

O- next group came back to the table without being called (procedures)
5:07 am

O- School nurse walked in...no disturbances. (proceudres)
5:14 am

S- made a big circle with notebooks and pencils ready to go. (procedures) T- "I'm impressed" responding to the making of the circle.
5:17 am

Critical Attributes

Proficient - The students are productively engaged during small group work.
Proficient - Transitions between large and small group activities are smooth.
Proficient - Routines for distribution and collection of materials and supplies work efficiently.
Proficient - Classroom routines function smoothly.

Summary

You clearly have your students and their materials organized in an effective way. Your procedures are clearly set and very little instructional time is wasted.

2d: Managing Student Behavior

[redacted]

Score: 3

Critical Attributes

Proficient - Standards of conduct appear to have been established.
Proficient - Student behavior is generally appropriate.
Proficient - The teacher frequently monitors student behavior.
Proficient - Teachers response to student misbehavior is effective.
Proficient - Teacher acknowledges good behavior

Summary

No notes to share because there was no student behavior problems during the lesson.

2e: Organizing physical space

[redacted]

Score: 3

Evidence

O- classroom neat and organized...space is used very well.
4:59 am

Critical Attributes

Proficient - The classroom is safe, and all students are able to see and hear.
Proficient - The classroom is arranged to support the instructional goals and learning activities.
Proficient - The teacher makes appropriate use of available technology.

Summary

Room and materials are organized and neat. The use of technology is evident.

3a: Communicating with students

[redacted]

Score: 4

Evidence

T- "next person can now share" S- sharing their for or against zoos.
4:46 am

T- "I need all eyes and ears" T- "We will discuss whole group later, right now we are going to do our Daily."
4:55 am

T- sat with a small group for the first Daily. Gave clear directions to what was expected. "Alright."
4:57 am

T-" As we read the two paragraphs I would like for you to think about cause and effect." "What do we think a cause is?" S- wrote answer down. T- "What is the effect?" S- wrote answer down.
5:01 am

T- "We have six details, we need to decide on the main idea." T-" talk to each other to see if you can come up with one sentence that will combine these."
5:13 am

T- "For our final mini-lesson, we need our notebook" "Let's see if we can do this in 2 minutes." "Let's mnake our big circle."
5:16 am

T- "It's not an argument, it's a discussion." (set the table for the group conversation)
5:19 am

T- setting up for the big debate "If you are for zoos raise your hand" Against?" Raise your hand." "no changing or this won't work"
5:31 am

Critical Attributes

Distinguished - In addition to the characteristics of proficient,
Distinguished - The teacher points out possible areas for misunderstanding.
Distinguished - Teacher explains content clearly and imaginatively, using metaphors and analogies to bring content to life.
Distinguished - All students seem to understand the presentation.
Distinguished - The teacher invites students to explain the content to the class, or to classmates.
Distinguished - Teacher uses rich language, offering brief vocabulary lessons where appropriate.

Summary

You clearly have skills in this area. Your students were able to clearly grasp the information needed to complete the assigned task. Communication between teacher and students is respectful.

3b: Using questioning and discussion techniques

[redacted] **Score: 4**

Evidence

S- shared their opinion...then students in the group were able to ask questions that they may have. (student to student)
4:48 am

S- "where would the big animals go?" S- "In the natural habitat." (respectfully answered the question)
4:54 am

T- "What is the main idea of this sentence?" S- responded T- "OK"
5:00 am

T-" As we read the two paragraphs I would like for you to think about cause and effect." "What do we think a cause is?" S- wrote answer down. T- "What is the effect?" S- wrote answer down.
5:01 am

T- "Any other animal or situation similar to that cause or effect?" (Q- deeper thinking, connection)
5:03 am

T- "What do you think" S- responded T- "way to go, I was thinking the same thing?"
5:12 am

T- "We have six details, we need to decide on the main idea." T-" talk to each other to see if you can come up with one sentence that will combine these."
5:13 am

O- this type of discussion leads to a better understanding of the debate. They did it in a respectful way. (questioning)
5:22 am

T-" Will you tell us about the analogy of the story that you read?" talking to a student who read a book recently. This sparked more conversation.
5:22 am

O- teacher continued to add questions to continue the conversation. (Questions were built off of the conversation from the students)
5:27 am

T- "Levi?" have you gone to our zoo?" Are you as guilty as I am for throwing corn at the animals?" S- responded with a smile. (this was in response to a student response to a question)
5:30 am

Critical Attributes

Distinguished - In addition to the characteristics of proficient,

Distinguished - Students initiate higher-order questions.
Distinguished - Students extend the discussion, enriching it.
Distinguished - Students invite comments from their classmates during a discussion.

Summary

It is evident that you have worked to improve this area. I observed your questioning strategies to be mostly "higher" level thinking. This is what we are striving for school-wide. The small group questioning from student to student was impressive.

3c: Engaging students in learning

[redacted] **Score: 3**

Evidence

T- "next person can now share" S- sharing their for or against zoos.
4:46 am

S- shared their opinion...then students in the group were able to ask questions that they may have. (student to student)
4:48 am

O- student sharing his thoughts about the zoo issue...other students listened and jotted down questions to ask at the end.
4:52 am

T-" As we read the two paragraphs I would like for you to think about cause and effect." "What do we think a cause is?" S- wrote answer down. T- "What is the effect?" S- wrote answer down.
5:01 am

T- "We have six details, we need to decide on the main idea." T-" talk to each other to see if you can come up with one sentence that will combine these."
5:13 am

T- setting up for the big debate "If you are for zoos raise your hand" Against?" Raise your hand." "no changing or this won't work"
5:31 am

Critical Attributes

Proficient - Most students are intellectually engaged in the lesson.
Proficient - Learning tasks have multiple correct responses or approaches and/or demand higher-order thinking
Proficient - Students have some choice in how they complete learning tasks.
Proficient - There is a mix of different types of groupings, suitable to the lesson objectives.

Proficient - Materials and resources support the learning goals and require intellectual engagement, as appropriate.
Proficient - The pacing of the lesson provides students the time needed to be intellectually engaged.

Summary

Student engagement in the lesson was evident. The student to student conversations made the lesson more enriching. Well done!

3d: Using assessment in instruction

[redacted]

Score: 3

Evidence

S- shared their opinion...then students in the group were able to ask questions that they may have. (student to student)
4:48 am

O- student sharing his thoughts about the zoo issue...other students listened and jotted down questions to ask at the end.
4:52 am

S- "where would the big animals go?" S- "In the natural habitat." (respectfully answered the question)
4:54 am

Critical Attributes

Proficient - Students indicate that they clearly understand the characteristics of high-quality work.
Proficient - The teacher elicits evidence of student understanding during the lesson Students are invited to assess their own work and make improvements.
Proficient - Feedback includes specific and timely guidance for at least groups of students
Proficient - The teacher attempts to engage students in self- or peer-assessment.
Proficient - When necessary, the teacher makes adjustments to the lesson to enhance understanding by groups of students.
Distinguished - Teacher makes frequent use of strategies to elicit information about individual student understanding.
Distinguished - Feedback to students is specific and timely, and is provided from many sources, including other students.
Distinguished - Students monitor their own understanding, either on their own initiative or as a result of tasks set by the teacher.

Summary

Your feedback to students was clear and concise. Your questioning strategies enable ou to understand and feel comfrotable knowing if your students understand the material.

3e: Demonstrating flexibility and responsiveness

[redacted]

Score: NA

Summary

No evidence to score.

Notes

[redacted]

Q- "Was this a typical group discussion?" (format)
5:24 am

Q- "Is it ok if all students don't share in the conversation?"
5:25 am

Summary

[redacted]

Recommendations:

Continue being a positive leader throughout our building. Continue using "new" ideas to be creative and inventive with your students.

Areas of Strength:

Clearly your ability to communicate and have enriching discussions with your students is a strength of yours. Your organization and procedures for your students is very noticeable. Your ability to connect with students is a skill that comes very naturally to you. Your focus on student growth is greatly appreciated and drives you to become a better instructor.

Areas for Growth:

Continue to use technology to enhance your instruction.

Additional Comments:

I enjoyed my time in your room. You have created a positive and productive learning environment. I appreciate what you have done for the "good" of the school. I know not all is "noticed" by everyone, but know that I greatly appreciate your efforts! Keep up the great work!!

Report exported on Aug 15, 2014 - 11:23 AM

Sample Value-Added Report for Principal

TLES
Teacher & Leader Evaluation Systems
A Study at American Institutes for Research

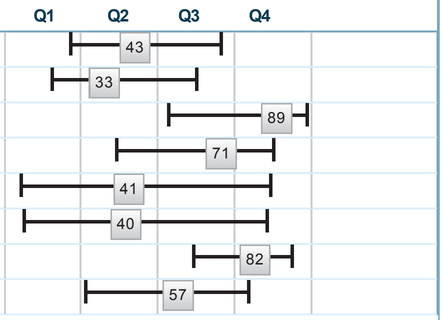
Value-Added Scores for Teachers in School 3

2010-2011/2011-2012

Legend: Quartile
■ Q1 ■ Q2 ■ Q3 ■ Q4

Name	Number of Student Scores	Number of Teachers	Value-Added Score with Standard Error	Average Teacher Value-Added Score with Standard Error	% Teachers at Each Quartile
District 1	55929	784	0.00±0.00	-0.01±0.08	24 25 25 26
School 3	311	7	0.02±0.03	0.03±0.09	43 29 29

Name	Number of Student Scores	Value-Added Score with Standard Error	Percentile for Value-Added Score with Confidence Range
Teacher A	66	-0.03±0.07	43
Teacher B	62	-0.06±0.07	33
Teacher C	22	0.17±0.14†	89
Teacher D	47	0.06±0.09	71
Teacher E	10	-0.03±0.15†	41
Teacher F	12	-0.04±0.14†	40
Teacher G	58	0.12±0.08	82
Teacher H	44	0.01±0.08	57



Based on data from 2010-2011/2011-2012

Value-added scores indicated with a † are based on single-year averages rather than two-year averages. Research has shown that value-added scores can vary substantially from one year to the next, and averaging over two years will help ensure that the reported scores reflect teaching effectiveness that persists over time, rather than year-to-year fluctuations in teaching effectiveness that may occur due to teachers' personal circumstances, reform initiatives, or fluctuations due to other factors (such as relatively small numbers of students in some classrooms). For teacher with only one-year scores, the standard errors may be larger (and it may be harder to distinguish the teacher's performance from average).

When there are fewer than ten student scores in a particular category, all columns other than Number of Student Scores will have asterisks. Reliable results cannot be generated from a small number of student scores.

TLES
Teacher & Leader Evaluation Systems
A Study at American Institutes for Research

Value-Added Scores for Teachers in Schoo 3 by Subject

2010-2011/2011-2012

Legend: Quartile
■ Q1 ■ Q2 ■ Q3 ■ Q4

Name	Subject	Number of Student Scores	Number of Teachers	Value-Added Score with Standard Error	Average Teacher Value-Added Score with Standard Error	% Teachers at Each Quartile
District 1	Overall	55929	784	0.00±0.00	-0.01±0.08	24 25 25 26
	Mathematics	27536	640	0.00±0.01	-0.01±0.11	24 25 25 26
	Reading	28393	642	0.00±0.00	0.00±0.10	24 25 25 26
School 3	Overall	311	7	0.02±0.03	0.03±0.09	43 29 29
	Mathematics	157	7	0.08±0.04	0.14±0.12	29 29 43
	Reading	154	7	-0.04±0.04	-0.06±0.11	71 29

Name	Subject	Number of Student Scores	Value-Added Score with Standard Error	Percentile for Value-Added Score with Confidence Range			
				Q1	Q2	Q3	Q4
Teacher A	Overall	66	-0.03±0.07		43		
	Mathematics	33	0.02±0.08			59	
	Reading	33	-0.08±0.08	23			
Teacher B	Overall	62	-0.06±0.07		33		
	Mathematics	31	-0.04±0.09		46		
	Reading	31	-0.09±0.08	21			
Teacher C	Overall	22	0.17±0.14†				89
	Mathematics	11	0.52±0.17†				99
	Reading	11	-0.15±0.16†	11			
Teacher D	Overall	47	0.06±0.09			71	
	Mathematics	23	0.26±0.11				90
	Reading	24	-0.11±0.10	17			
Teacher E	Overall	10	-0.03±0.15†		41		
	Mathematics	5	*†				
	Reading	5	*†				
Teacher F	Overall	12	-0.04±0.14†		40		
	Mathematics	6	*†				
	Reading	6	*†				

Teacher G	Overall	58	0.12±0.08	82
	Mathematics	29	0.17±0.10	81
	Reading	29	0.08±0.09	77
Teacher H	Overall	44	0.01±0.08	57
	Mathematics	24	-0.07±0.10	38
	Reading	20	0.08±0.09	79

Based on data from 2010-2011/2011-2012
Report Generated: 2/23/2014 8:58:34 PM EST

Value-added scores indicated with a † are based on single-year averages rather than two-year averages. Research has shown that value-added scores can vary substantially from one year to the next, and averaging over two years will help ensure that the reported scores reflect teaching effectiveness that persists over time, rather than year-to-year fluctuations in teaching effectiveness that may occur due to teachers' personal circumstances, reform initiatives, or fluctuations due to other factors (such as relatively small numbers of students in some classrooms). For teacher with only one-year scores, the standard errors may be larger (and it may be harder to distinguish the teacher's performance from average).

When there are fewer than ten student scores in a particular category, all columns other than Number of Student Scores will have asterisks. Reliable results cannot be generated from a small number of student scores.

TLES
Teacher & Leader Evaluation Systems
A Study at American Institutes for Research

Value-Added Scores for Teachers in School 3 by Grade and Subject

2010-2011/2011-2012

Comparison Scores

Name	Subject/Grade	Number of Student Scores	Number of Teachers	Value-Added Score with Standard Error	Average Teacher Value-Added Score with Standard Error
District 1	All - Grade 4	10145	277	0.00±0.01	-0.01±0.11
	All - Grade 5	11943	230	0.00±0.01	-0.01±0.08
	All - Grade 6	11664	144	-0.01±0.01	-0.01±0.09
	All - Grade 7	10848	135	0.00±0.01	0.00±0.07
	All - Grade 8	11329	142	0.00±0.01	0.00±0.07
	Mathematics - Grade 4	5058	274	-0.01±0.01	-0.02±0.14
	Mathematics - Grade 5	5945	219	0.00±0.01	-0.01±0.10
	Mathematics - Grade 6	5529	77	-0.02±0.03	-0.01±0.10
	Mathematics - Grade 7	5371	70	-0.01±0.02	-0.01±0.08
	Mathematics - Grade 8	5633	77	0.00±0.02	-0.01±0.10
	Reading - Grade 4	5087	275	0.00±0.01	0.00±0.13
	Reading - Grade 5	5998	224	0.00±0.01	-0.01±0.10
	Reading - Grade 6	6135	84	-0.01±0.01	0.00±0.09
	Reading - Grade 7	5477	68	0.00±0.01	0.00±0.05
	Reading - Grade 8	5696	73	0.00±0.01	0.00±0.06
School 3	All - Grade 4	139	4	0.08±0.05	0.08±0.11
	All - Grade 5	172	3	-0.03±0.04	-0.03±0.07
	Mathematics - Grade 4	69	4	0.24±0.06	0.26±0.14
	Mathematics - Grade 5	88	3	-0.02±0.05	-0.03±0.09
	Reading - Grade 4	70	4	-0.05±0.06	-0.08±0.13
	Reading - Grade 5	84	3	-0.04±0.05	-0.03±0.09

Name	Subject/Grade	Number of Student Scores	Value-Added Score with Standard Error
Teacher A	All - Grade 5	66	-0.03±0.07
	Mathematics - Grade 5	33	0.02±0.08
	Reading - Grade 5	33	-0.08±0.08
Teacher B	All - Grade 5	62	-0.06±0.07
	Mathematics - Grade 5	31	-0.04±0.09
	Reading - Grade 5	31	-0.09±0.08
Teacher C	All - Grade 4	22	0.17±0.14†
	Mathematics - Grade 4	11	0.52±0.17†
	Reading - Grade 4	11	-0.15±0.16†
Teacher D	All - Grade 4	47	0.06±0.09
	Mathematics - Grade 4	23	0.26±0.11
	Reading - Grade 4	24	-0.11±0.10
Teacher E	All - Grade 4	10	-0.03±0.15†
	Mathematics - Grade 4	5	*
	Reading - Grade 4	5	*
Teacher F	All - Grade 4	12	-0.04±0.14†
	Mathematics - Grade 4	6	*
	Reading - Grade 4	6	*
Teacher G	All - Grade 4	58	0.12±0.08
	Mathematics - Grade 4	29	0.17±0.10
	Reading - Grade 4	29	0.08±0.09
Teacher H	All - Grade 5	44	0.01±0.08
	Mathematics - Grade 5	24	-0.07±0.10
	Reading - Grade 5	20	0.08±0.09

Based on data from 2010-2011/2011-2012

Value-added scores indicated with a † are based on single-year averages rather than two-year averages. Research has shown that value-added scores can vary substantially from one year to the next, and averaging over two years will help ensure that the reported scores reflect teaching effectiveness that persists over time, rather than year-to-year fluctuations in teaching effectiveness that may occur due to teachers' personal circumstances, reform initiatives, or fluctuations due to other factors (such as relatively small numbers of students in some classrooms). For teacher with only one-year scores, the standard errors may be larger (and it may be harder to distinguish the teacher's performance from average).

When there are fewer than ten student scores in a particular category, all columns other than Number of Student Scores will have asterisks. Reliable results cannot be generated from a small number of student scores.