

U.S. Department of Education
December 2017

The Impact of Providing Performance Feedback to Teachers and Principals

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

Melanie Ali
Project Officer
Institute of Education Sciences

This page has been left blank for double-sided copying

The Impact of Providing Performance Feedback to Teachers and Principals

December 2017

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

Melanie Ali
Project Officer
Institute of Education Sciences

NCEE 2018-4001
U.S. DEPARTMENT OF EDUCATION



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

U.S. Department of Education
Betsy DeVos
Secretary

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

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

December 2017

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

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

This report is in the public domain. Authorization to reproduce it in whole or in part is granted. While permission to reprint this publication is not necessary, the citation should be:

Garet, M.S., Wayne, A.J., Brown, S., Rickles, J., Song, M., and Manzeske, D. (2017). *The Impact of Providing Performance Feedback to Teachers and Principals* (NCEE 2018-4001). 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, Nicol Christie, Debbie Davidson-Gibbs, Mark Garabaldi, Jessica Milton, Elaine Liebesman, Amy Potemski, Roshni Menon, Marian Eaton, Gur Hoshen, Zhongjie Sun, and Michele Cadigan. Additional AIR staff worked tirelessly on data collection: Dorothy Seidel, Lauren Staley, Cheryl Pruce, Sarah Bardack, Makeda Amelga, and Lindsey Mitchell. For their efforts to identify the partner districts, we thank the recruitment leaders Anja Kurki and Rebecca Herman and the team of senior recruiters: Kirk Walters, James Taylor, Marlene Darwin, Nicholas Sorensen, Mark Garibaldi, Carmen Martinez, Nicol Christie, Kathleen Perez-Lopez, and Emily Rosenthal. The study authors are also grateful to Rachel Garrett, Jenifer Harr-Robins, Luke Keele, and Paul Bailey for their help with data analyses, in addition to Connie Conroy who provided administrative assistance throughout the project. We also thank Jeanette Moses for her meticulous work supporting the development of the study reports, making sure they were ready for each stage of review and publication.

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

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

Disclosure of Potential Conflicts of Interest

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

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

Contents

	Page
Acknowledgments.....	vii
Disclosure of Potential Conflicts of Interest.....	ix
Executive Summary	ES-1
Study Overview	ES-2
Implementation of the Intervention	ES-5
Contrast in Educators' Experience of Feedback	ES-9
Impact on Classroom Practice, Principal Leadership, and Student Achievement	ES-11
Association Among Classroom Practice, Leadership, and Achievement.....	ES-16
Chapter 1. Introduction	1
Overview of the Intervention	2
Theory of Action and Research Questions	5
Overview of Study Design.....	8
Chapter 2. Implementation of the Performance Measures and Feedback	21
The Intervention's Measures of Teacher Classroom Practice	22
The Intervention's Measure of Student Growth	33
The Intervention's Measure of Principal Leadership	41
Summary	47
Chapter 3. Impact of Performance Feedback on Teacher, Principal, and Student Outcomes	49
Contrast in Educators' Experience of Feedback	51
Impact on Initial Outcomes.....	55
Impact on Classroom Practice, Principal Leadership, and Student Achievement	63
Association Among Classroom Practice, Leadership, and Achievement.....	77
Summary	78
References.....	79
Appendix A. Details About the Study Sample	A-1
Appendix B. Details About Data Collection and Outcome Measures.....	B-1
Appendix C. Technical Details About Reliability Estimation.....	C-1
Appendix D. Supplemental Findings About the Implementation of the Intervention's Measures of Classroom Practice	D-1

Appendix E. Technical Details About the Estimation of Value-Added Scores	E-1
Appendix F. Supplemental Findings About the Implementation of the Intervention’s Measure of Student Growth	F-1
Appendix G. Supplemental Findings About the Implementation of the Intervention’s Measure of Principal Leadership	G-1
Appendix H. Technical Details About Analyses Assessing Treatment-Control Differences in Educators’ Experiences and Impacts on Outcomes	H-1
Appendix I. Supporting Exhibits for Analyses of Educators’ Experiences and Initial Outcomes	I-1
Appendix J. Supporting Exhibits for Impact Analyses.....	J-1
Appendix K. Sample Reports	K-1

List of Exhibits

	Page
Exhibit ES.1. Number of feedback sessions with ratings and written narrative and duration of oral feedback that an average teacher reported receiving, by treatment status and year.....	ES-10
Exhibit ES.2. Number of feedback instances and duration of oral feedback that principals reported receiving, by treatment status and year	ES-11
Exhibit ES.3. Average CLASS and FFT scores, based on coding of video-recorded lessons by study team, by treatment status, Year 2.....	ES-13
Exhibit ES.4. Average rating of principal instructional leadership and teacher-principal trust, by treatment status and year.....	ES-14
Exhibit ES.5. Average reading/English language arts and mathematics achievement, by treatment status and year.....	ES-15
Exhibit 1.1. Theory of action	6
Exhibit 1.2. District selection and recruitment process	10
Exhibit 1.3. Characteristics of all districts in the United States and districts that participated in the study	11
Exhibit 1.4. Policies and practices for performance feedback to teachers, by district	13
Exhibit 1.5. Policies and practices for performance feedback to principals, by district.....	14
Exhibit 1.6. Random assignment results, fall 2012	15
Exhibit 2.1. Domains and dimensions of classroom practice for CLASS and FFT	23
Exhibit 2.2. Mean number of feedback sessions treatment teachers received in each study year and in total.....	25
Exhibit 2.3. Distribution of teachers across CLASS and FFT performance levels for Windows 1 and 4 and for the 4-Window average, Year 2.....	27
Exhibit 2.4. Percentage of treatment teachers and principals who agreed somewhat or strongly with each statement about the feedback they received from the study's CLASS/FFT observations, Year 2	32
Exhibit 2.5. Timeline for estimating value-added scores and delivering student growth reports	33
Exhibit 2.6. Distribution of treatment teachers based on whether their value-added score in each wave was considered measurably above or below the district average, overall and by subject.....	38
Exhibit 2.7. Percentage of treatment teachers and principals who agreed somewhat or strongly with statements about the student growth reports they viewed, Year 2	40
Exhibit 2.8. VAL-ED core components and key processes.....	41
Exhibit 2.9. Distribution of treatment principals across performance levels based on VAL-ED overall scores in fall and spring, by year	43

Exhibit 2.10. Correlations between VAL-ED respondent group overall scores from different respondent groups in fall and spring, by year	45
Exhibit 2.11. Percentage of treatment principals who agreed somewhat or strongly with statements about the feedback they received from the VAL-ED, Year 2.....	47
Exhibit 3.1. Number of feedback sessions with ratings and written narrative and duration of oral feedback that an average teacher reported receiving, by treatment status and year.....	52
Exhibit 3.2. Percentage of teachers who reported receiving specific types of student achievement information, by treatment status and year.....	54
Exhibit 3.3. Number of feedback instances and duration of oral feedback that principals reported receiving, by treatment status and year	55
Exhibit 3.4. Percentage of teachers reporting that they discussed, were interested in improving, and participated in professional development covering at least one area of practice measured by the CLASS or FFT or at least one area not measured, by treatment status, Year 2	57
Exhibit 3.5. Teachers' self-ratings of their effectiveness in boosting students' reading/ELA and mathematics achievement, by treatment status and year.....	59
Exhibit 3.6. Percentage of principals reporting that they discussed, were interested in improving, and participated in professional development covering at least one area of practice measured by the VAL-ED, by treatment status, Year 2.....	61
Exhibit 3.7. Principals' self-rating of their effectiveness in instructional leadership and other forms of leadership, by treatment status, Year 2	63
Exhibit 3.8. Average CLASS and FFT scores, based on coding of video-recorded lessons by study team, by treatment status, Year 2	66
Exhibit 3.9. Average CLASS and FFT scores in CLASS districts and FFT districts, based on coding of video-recorded lessons by study team, by treatment status, Year 2.....	68
Exhibit 3.10. Average rating of principal instructional leadership and teacher-principal trust, by treatment status and year.....	70
Exhibit 3.11. Average rating of principal instructional leadership and teacher-principal trust in CLASS districts and FFT districts, by treatment status and year.....	72
Exhibit 3.12. Average reading/ELA and mathematics achievement, by treatment status and year.....	74
Exhibit 3.13. Average reading/ELA and mathematics achievement in CLASS districts and FFT districts, by treatment status and year	76

Exhibit A.1. Background characteristics of elementary schools in the study sample, elementary schools in similarly sized districts, and the national population, baseline year	A-1
Exhibit A.2. Background characteristics for middle schools in the study sample, middle schools in similarly sized districts, and the national population, baseline year.....	A-2
Exhibit A.3. Background characteristics for schools in CLASS and FFT districts, baseline year.....	A-3
Exhibit A.4a. School background characteristics, by treatment status, baseline year	A-4
Exhibit A.4b. School background characteristics in CLASS districts, by treatment status, baseline year.....	A-4
Exhibit A.4c. School background characteristics in FFT districts, by treatment status, baseline year.....	A-5
Exhibit A.4d. Principal background characteristics, by treatment status, fall of Year 1	A-5
Exhibit A.4e. Background characteristics of principals in CLASS districts, by treatment status, fall of Year 1	A-6
Exhibit A.4f. Background characteristics of principals in FFT districts, by treatment status, fall of Year 1	A-6
Exhibit A.4g. Teacher background characteristics, by treatment status, fall of Year 1.....	A-7
Exhibit A.4h. Background characteristics of teachers in CLASS districts, by treatment status, fall of Year 1	A-7
Exhibit A.4i. Background characteristics of teachers in FFT districts, by treatment status, fall of Year 1	A-8
Exhibit A.4j. Student background characteristics, by treatment status, baseline year.....	A-8
Exhibit A.5. Principal turnover across study years.....	A-9
Exhibit A.6. Teacher turnover across study years	A-10
Exhibit A.7. Student turnover across study years, reading/ELA achievement impact sample	A-11
Exhibit A.8. Student turnover across study years, mathematics achievement impact sample	A-12
Exhibit A.9. Percentage of principals, teachers, and students who exited between spring Year 1 and 2, by treatment status.....	A-13
Exhibit A.10. Realized minimum detectable effect sizes for educator and student outcomes, by year	A-14
Exhibit B.1. Data collection schedule for intervention implementation data in each study year.....	B-1
Exhibit B.2 Response rates for teacher survey, principal survey, and video-recording, overall and by treatment status	B-3
Exhibit B.3. Item composition and reliabilities of principal leadership scales.....	B-6

Exhibit C.1. Summary of reliability estimates for measures of educator performance	C-2
Exhibit C.2. Estimated reliabilities for CLASS overall scores and dimension scores, Year 2	C-7
Exhibit C.3. Estimated reliabilities for FFT overall scores and dimension scores, Year 2	C-8
Exhibit C.4. Estimated variance components and reliabilities for dimension score differences for CLASS and FFT, Year 2	C-10
Exhibit C.5. Estimated reliabilities for value-added scores based on two years of student growth data, Year 2	C-12
Exhibit C.6. Estimated variance components and reliability for subject-specific value- added score differences, Year 2	C-13
Exhibit C.7. Estimated reliabilities for VAL-ED overall scores and dimension scores, fall of Year 2	C-15
Exhibit C.8. Estimated reliabilities for VAL-ED overall scores and dimension scores, spring of Year 2	C-16
Exhibit C.9. Estimated variance components and reliabilities for VAL-ED dimension score differences, fall of Year 2	C-17
Exhibit C.10. Estimated variance components and reliabilities for VAL-ED dimension score differences, spring of Year 2	C-18
Exhibit D.1. Percentage of treatment principals who agreed somewhat or strongly with each statement about the observations they conducted, by year	D-1
Exhibit D.2. Mean number of feedback sessions K-3 treatment teachers received, by year and in total	D-1
Exhibit D.3. Percentage of study-hired observers who reported that they engaged in a given activity in two-thirds or more of the feedback sessions they conducted, by year	D-2
Exhibit D.4. Percentage of study-hired observers who reported that teachers were engaged in the discussion in two-thirds or more of the feedback sessions they conducted, by year	D-2
Exhibit D.5. Average percentage of teachers that study-hired observers felt needed significant or some help according to the CLASS or FFT instrument, Year 2	D-2
Exhibit D.6a. Distribution of K-3 teachers across performance levels based on CLASS overall scores in each observation window, and the two-window average in each year	D-3
Exhibit D.6b. Distribution of K-3 teachers across performance levels based on FFT overall scores in each observation window, and the two-window average in each year	D-3
Exhibit D.7a. Distribution of teachers based on their CLASS overall scores in each observation window and the four-window average, by year	D-4

Exhibit D.7b. Distribution of teachers based on their FFT overall scores in each observation window and the four-window average, by year	D-5
Exhibit D.7c. Descriptive statistics for CLASS and FFT overall scores in each observation window, by year	D-6
Exhibit D.7d. Distribution of CLASS overall scores based on video-recorded lessons for treatment teachers in CLASS districts, and FFT scores for treatment teachers in FFT districts, spring Year 2	D-7
Exhibit D.7e. Pairwise correlations of intervention observation scores and video-recorded lesson scores with prior-year value-added, for treatment teachers in CLASS and FFT districts, Year 2	D-8
Exhibit D.8. Descriptive statistics for average CLASS and FFT observation scores in each year, by observer type.....	D-9
Exhibit D.9a. Descriptive statistics for four-window average CLASS observation scores and two-round average video-recorded lesson scores, for treatment teachers in CLASS districts, by domain and dimension, Year 2	D-10
Exhibit D.9b. Descriptive statistics for four-window average FFT observation scores and two-round average video-recorded lesson scores, for treatment teachers in FFT districts, by domain and dimension, Year 2	D-11
Exhibit D.10. Percentage of teachers whose dimension scores spanned one, two, three, or four performance levels, by observation window.....	D-12
Exhibit D.11. Percentage of treatment teachers who agreed somewhat or strongly with each statement about the feedback they received from the study's CLASS/FFT observations, compared with the feedback received prior to the intervention as part of their district's approach to formal evaluation, Year 2	D-13
Exhibit D.12. Percentage of principals who agreed somewhat or strongly with each statement about the fairness and validity of CLASS or FFT, Year 2 ..	D-13
Exhibit F.1. Percentage of treatment teachers with sufficient data to estimate value-added scores, and percentage of teachers whose scores were based on two years of data, by year	F-1
Exhibit F.2. Percentage Distribution of treatment teachers based on whether their subject area value-added scores were measurably above or below the district average, by year	F-1
Exhibit F.3. Percentage of treatment teachers who agreed somewhat or strongly with statements about the student growth reports they viewed, Year 2.....	F-2
Exhibit F.4. Percentage of treatment principals who agreed somewhat or strongly with statements about the student growth reports they viewed, Year 2.....	F-2
Exhibit G.1. Definitions of VAL-ED core components and key processes.....	G-1
Exhibit G.2. Sample VAL-ED survey items.....	G-2
Exhibit G.3. Results overview from a sample VAL-ED report.....	G-3

Exhibit G.4. Results by respondent group from a sample VAL-ED report	G-4
Exhibit G.5. Summary of component-by-process scores from a sample VAL-ED report	G-5
Exhibit G.6. Descriptive statistics for average VAL-ED overall scores in fall and spring of each year	G-6
Exhibit G.7. Descriptive statistics for average VAL-ED overall scores in fall and spring of each year, by respondent group	G-6
Exhibit G.8. Percentage of principals whose VAL-ED scores spanned one, two, three, or four performance levels, by wave	G-7
Exhibit G.9. Percentage of treatment principals who agreed somewhat or strongly with statements about the feedback they received from the VAL-ED, Year 2	G-7
Exhibit I.1a. Percentage of teachers who reported receiving ratings on their classroom practice, being observed by their principal, and being observed by someone from outside of their school, overall and within CLASS and FFT districts, by treatment status, Year 1	I-1
Exhibit I.1b. Percentage of teachers who reported receiving ratings on their classroom practice, being observed by their principal, and being observed by someone from outside of their school, overall and within CLASS and FFT districts, by treatment status, Year 2	I-2
Exhibit I.2a. Number of feedback instances and duration of feedback that an average teacher reported receiving, overall and within CLASS and FFT districts, by treatment status, Year 1	I-3
Exhibit I.2b. Number of feedback instances and duration of feedback that an average teacher reported receiving, overall and within CLASS and FFT districts, by treatment status, Year 2	I-4
Exhibit I.3a. Percentage of teachers who reported receiving specific types of student achievement information, overall and within CLASS and FFT districts, by treatment status, Year 1	I-5
Exhibit I.3b. Percentage of teachers who reported receiving specific types of student achievement information, overall and within CLASS and FFT districts, by treatment status, Year 2	I-6
Exhibit I.4a. Number of feedback instances and duration of feedback that an average principal reported receiving, overall and within CLASS and FFT districts, by treatment status, Year 1	I-7
Exhibit I.4b. Number of feedback instances and duration of feedback that an average principal reported receiving, overall and within CLASS and FFT districts, by treatment status, Year 2	I-8
Exhibit I.5a. Percentage of teachers who reported discussing areas of practice related to CLASS/FFT with someone who provided them with feedback during the school year, overall and within CLASS and FFT districts, by treatment status, Year 1	I-9

Exhibit I.5b. Percentage of teachers who reported discussing areas of practice related to CLASS/FFT with someone who provided them with feedback during the school year, overall and within CLASS and FFT districts, by treatment status, Year 2	I-10
Exhibit I.6a. Percentage of teachers who reported discussing areas of practice not related to CLASS/FFT with someone who provided them with feedback during the school year, overall and within CLASS and FFT districts, by treatment status, Year 1	I-11
Exhibit I.6b. Percentage of teachers who reported discussing areas of practice not related to CLASS/FFT with someone who provided them with feedback during the school year, overall and within CLASS and FFT districts, by treatment status, Year 2	I-12
Exhibit I.7a. Percentage of teachers who reported wanting to improve in areas of practice related to CLASS/FFT by a moderate or large amount, overall and within CLASS and FFT districts, by treatment status, Year 1	I-13
Exhibit I.7b. Percentage of teachers who reported wanting to improve in areas of practice related to CLASS/FFT by a moderate or large amount, overall and within CLASS and FFT districts, by treatment status, Year 2	I-14
Exhibit I.8a. Percentage of teachers who reported wanting to improve in areas of practice not related to CLASS/FFT by a moderate or large amount, overall and within CLASS and FFT districts, by treatment status, Year 1	I-15
Exhibit I.8b. Percentage of teachers who reported wanting to improve in areas of practice not related to CLASS/FFT by a moderate or large amount, overall and within CLASS and FFT districts, by treatment status, Year 2	I-16
Exhibit I.9a. Percentage of teachers who reported that their professional development activities during Year 1 covered areas of practice related to CLASS/FFT to a moderate or large extent, overall and within CLASS and FFT districts, by treatment status.....	I-17
Exhibit I.9b. Percentage of teachers who reported that their professional development activities during the summer between Years 1 and 2 covered areas of practice related to CLASS/FFT to a moderate or large extent, overall and within CLASS and FFT districts, by treatment status	I-18
Exhibit I.9c. Percentage of teachers who reported that their professional development activities during Year 2 covered areas of practice related to CLASS/FFT to a moderate or large extent, overall and within CLASS and FFT districts, by treatment status.....	I-19
Exhibit I.10a. Percentage of teachers who reported that their professional development activities during Year 1 covered areas of practice not related to CLASS/FFT to a moderate or large extent, overall and within CLASS and FFT districts, by treatment status.....	I-20

Exhibit I.10b. Percentage of teachers who reported that their professional development activities during the summer between Years 1 and 2 covered areas of practice not related to CLASS/FFT to a moderate or large extent, overall and within CLASS and FFT districts, by treatment status	I-21
Exhibit I.10c. Percentage of teachers who reported that their professional development activities during Year 2 covered areas of practice not related to CLASS/FFT to a moderate or large extent, overall and within CLASS and FFT districts, by treatment status.....	I-22
Exhibit I.11. Teachers' self-appraisal of their effectiveness in boosting students' reading/ELA and mathematics achievement, overall and within CLASS and FFT districts, by treatment status and year.....	I-23
Exhibit I.12a. The association between teachers' self-appraisal of their effectiveness in boosting students' reading/ELA and mathematics achievement and their prior-value-added score, overall and within CLASS and FFT districts, by treatment status and year	I-24
Exhibit I.12b. Teachers' prior value-added percentile for teachers with self-appraisals in different categories, by subject, Year 1	I-25
Exhibit I.12c. Teachers' prior value-added percentile for teachers with self-appraisals in different categories, by subject, Year 2	I-26
Exhibit I.13a. Percentage of principals who reported discussing areas of practice related to the VAL-ED with a supervisor during the school year, overall and within CLASS and FFT districts, by treatment status, Year 1	I-27
Exhibit I.13b. Percentage of principals who reported discussing areas of practice related to the VAL-ED with a supervisor during the school year, overall and within CLASS and FFT districts, by treatment status, Year 2.....	I-28
Exhibit I.14a. Percentage of principals who reported discussing areas of practice not related to the VAL-ED with a supervisor during the school year, overall and within CLASS and FFT districts, by treatment status, Year 1	I-29
Exhibit I.14b. Percentage of principals who reported discussing areas of practice not related to the VAL-ED with a supervisor during the school year, overall and within CLASS and FFT districts, by treatment status, Year 2.....	I-30
Exhibit I.15a. Percentage of principals who reported wanting to improve in areas of practice related to the VAL-ED by a moderate or large amount, overall and within CLASS and FFT districts, by treatment status, Year 1	I-31
Exhibit I.15b. Percentage of principals who reported wanting to improve in areas of practice related to the VAL-ED by a moderate or large amount, overall and within CLASS and FFT districts, by treatment status, Year 2	I-32

Exhibit I.16a. Percentage of principals who reported wanting to improve in areas of practice not related to the VAL-ED by a moderate or large amount, overall and within CLASS and FFT districts, by treatment status, Year 1	I-33
Exhibit I.16b. Percentage of principals who reported wanting to improve in areas not related to the VAL-ED by a moderate or large amount, overall and within CLASS and FFT districts, by treatment status, Year 2.....	I-34
Exhibit I.17a. Percentage of principals who reported that their professional development activities during Year 1 covered areas related to the VAL-ED to a moderate or large extent, overall and within CLASS and FFT districts, by treatment status.....	I-35
Exhibit I.17b. Percentage of principals who reported that their professional development activities during the summer between Years 1 and 2 covered areas related to the VAL-ED to a moderate or large extent, overall and within CLASS and FFT districts, by treatment status	I-36
Exhibit I.17c. Percentage of principals who reported that their professional development activities during Year 2 covered areas related to the VAL-ED to a moderate or large extent, overall and within CLASS and FFT districts, by treatment status.....	I-37
Exhibit I.18a. Percentage of principals who reported that their professional development activities during Year 1 covered areas of practice not related to the VAL-ED to a moderate or large extent, overall and within CLASS and FFT districts, by treatment status	I-38
Exhibit I.18b. Percentage of principals who reported that their professional development activities during the summer between Year 1 and 2 covered areas of practice not related to the VAL-ED to a moderate or large extent, overall and within CLASS and FFT districts, by treatment status	I-39
Exhibit I.18c. Percentage of principals who reported that their professional development activities during Year 2 covered areas not related to the VAL-ED to a moderate or large extent, overall and within CLASS and FFT districts, by treatment status.....	I-40
Exhibit I.19. Principals' self-appraisal of their effectiveness in instructional leadership and other forms of leadership, overall and within CLASS and FFT districts, overall and within CLASS and FFT districts, by treatment status, Year 2.....	I-41
Exhibit J.1. Background characteristics of teachers in the Year 2 teacher practice impact sample, overall and within CLASS and FFT districts, by treatment status	J-1
Exhibit J.2. Background characteristics of teachers in the Year 1 principal leadership impact sample, overall and within CLASS and FFT districts, by treatment status	J-2

Exhibit J.3. Background characteristics of teachers in the Year 2 principal leadership impact sample, overall and within CLASS and FFT districts, by treatment status	J-3
Exhibit J.4. Background characteristics of principals in the Year 1 principal leadership impact sample, overall and within CLASS and FFT districts, by treatment status	J-4
Exhibit J.5. Background characteristics of principals in the Year 2 principal leadership impact sample, overall and within CLASS and FFT districts, by treatment status	J-5
Exhibit J.6. Background characteristics of students in Year 1 reading/ELA achievement impact sample, by treatment status	J-6
Exhibit J.7. Background characteristics of students in Year 1 reading/ELA achievement impact sample in CLASS and FFT districts, by treatment status	J-7
Exhibit J.8. Background characteristics of students in Year 1 mathematics achievement impact sample, by treatment status	J-8
Exhibit J.9. Background characteristics of students in Year 1 mathematics achievement impact sample in CLASS and FFT districts, by treatment status	J-9
Exhibit J.10. Background characteristics of students in Year 2 reading/ELA achievement impact sample, by treatment status	J-10
Exhibit J.11. Background characteristics of students in Year 2 reading/ELA achievement impact sample in CLASS and FFT districts, by treatment status	J-11
Exhibit J.12. Background characteristics of students in Year 2 mathematics achievement impact sample, by treatment status	J-12
Exhibit J.13. Background characteristics of students in Year 2 mathematics achievement impact sample in CLASS and FFT districts, by treatment status	J-13
Exhibit J.14. Average CLASS and FFT overall scores, based on coding of video-recorded lessons by study team, overall and within CLASS and FFT districts, by treatment status, Year 2	J-14
Exhibit J.15. Average CLASS and FFT domain scores, based on coding of video-recorded lessons by study team, by treatment status, Year 2	J-14
Exhibit J.16. Average CLASS and FFT domain scores in CLASS and FFT districts, based on coding of video-recorded lessons by study team, by treatment status, Year 2	J-15
Exhibit J.17. Average CLASS and FFT overall scores, based on coding of video-recorded lessons by study team, by treatment status and district, Year 2	J-16
Exhibit J.18. Average CLASS and FFT overall scores without covariate adjustment, based on coding of video-recorded lessons by study team, overall and within CLASS and FFT districts, by treatment status, Year 2	J-17
Exhibit J.19. Average CLASS and FFT overall scores, based on coding of video-recorded lessons by study team, overall and within CLASS and FFT districts (excluding District 3), by treatment status, Year 2	J-17

Exhibit J.20. Average ratings of principal instructional leadership and teacher-principal trust, by treatment status, and year.....	J–18
Exhibit J.21. Average ratings of principal instructional leadership and teacher-principal trust in CLASS and FFT districts, by treatment status and year.....	J–18
Exhibit J.22. Average rating of principal instructional leadership, by treatment status, district, and year.....	J–19
Exhibit J.23. Average rating of teacher-principal trust, by treatment status, district, and year.....	J–20
Exhibit J.24. Average ratings of principal instructional leadership and teacher-principal trust without covariate adjustment, by treatment status and year	J–21
Exhibit J.25. Average ratings of principal instructional leadership and teacher-principal trust without covariate adjustment in CLASS and FFT districts, by treatment status and year.....	J–21
Exhibit J.26. Average reading/ELA and mathematics achievement, by treatment status and year.....	J–22
Exhibit J.27. Average reading/ELA and mathematics achievement in CLASS and FFT districts, by treatment status and year	J–23
Exhibit J.28. Average reading/ELA achievement, by treatment status, district, and year.....	J–24
Exhibit J.29. Average mathematics achievement, by treatment status, district, and year	J–25
Exhibit J.30. Average reading/ELA and mathematics achievement without covariate adjustment, by treatment status and year	J–26
Exhibit J.31. Average reading/ELA and mathematics achievement without covariate adjustment in CLASS and FFT districts, by treatment status and year.....	J–27
Exhibit J.32. Average reading/ELA and mathematics achievement adjusted for prior achievement in both reading/ELA and mathematics, by treatment status and year	J–28
Exhibit J.33. Average reading/ELA and mathematics achievement adjusted for prior achievement in both reading/ELA and mathematics in CLASS and FFT districts, by treatment status and year	J–29
Exhibit J.34. Differential impact of intervention on CLASS and FFT overall scores for teachers with different probationary status, teachers with different prior value-added scores, and teachers at different school levels, Year 2.....	J–30
Exhibit J.35. Differential impact of intervention on principal instructional leadership and teacher-principal trust for middle school principals and elementary school principals, by year	J–30
Exhibit J.36. Differential impact of intervention on student achievement in reading/ELA and mathematics, for teachers with different probationary status, teachers with different prior value-added, and teachers at different school levels, by year.....	J–31

Exhibit J.37a. Estimated relationships between classroom practice, principal leadership, and student achievement in reading/ELA, by year	J-32
Exhibit J.37b. Estimated relationships between classroom practice, principal leadership, and student achievement in reading/ELA in CLASS districts, by year.....	J-32
Exhibit J.37c. Estimated relationships between classroom practice, principal leadership, and student achievement in reading/ELA in FFT districts, by year	J-33
Exhibit J.37d. Estimated relationships between classroom practice, principal leadership, and student achievement in mathematics, by year.....	J-33
Exhibit J.37e. Estimated relationships between classroom practice, principal leadership, and student achievement in mathematics in CLASS districts, by year.....	J-34
Exhibit J.37f. Estimated relationships between classroom practice, principal leadership, and student achievement in mathematics in FFT districts, by year	J-34

Executive Summary

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

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

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

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

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

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

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

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

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

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

⁵ See Wayne et al. (2016).

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

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

Study Overview

The study addressed five research questions:

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

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

Study Design

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

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

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

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

Data Sources

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

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

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

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

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

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

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

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

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

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

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

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

Analyses

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

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

Implementation of the Intervention

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Contrast in Educators’ Experience of Feedback

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

Did the intervention increase feedback for teachers?

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

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

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

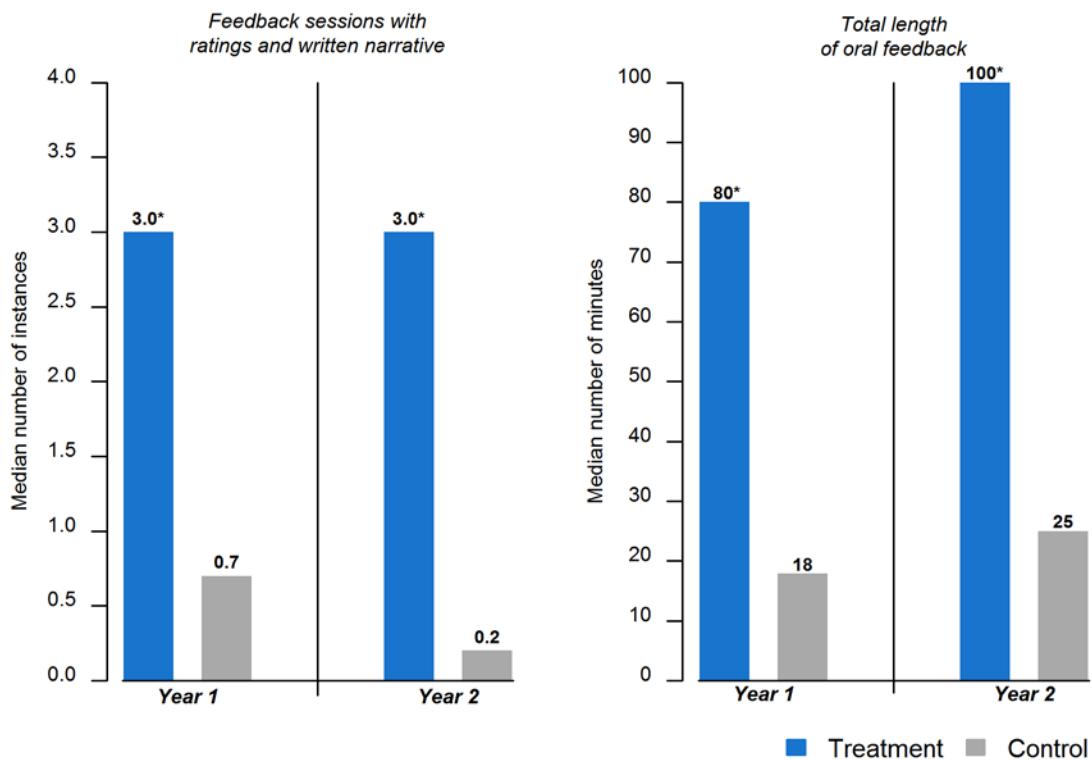


EXHIBIT READS: The average treatment teacher in Year 1 reported 3.0 feedback sessions with ratings and written narrative, compared with 0.7 for control teachers.

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

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

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

SOURCES: Spring 2013 and Spring 2014 Teacher Surveys.

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

Did the intervention increase feedback for principals?

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

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

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

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

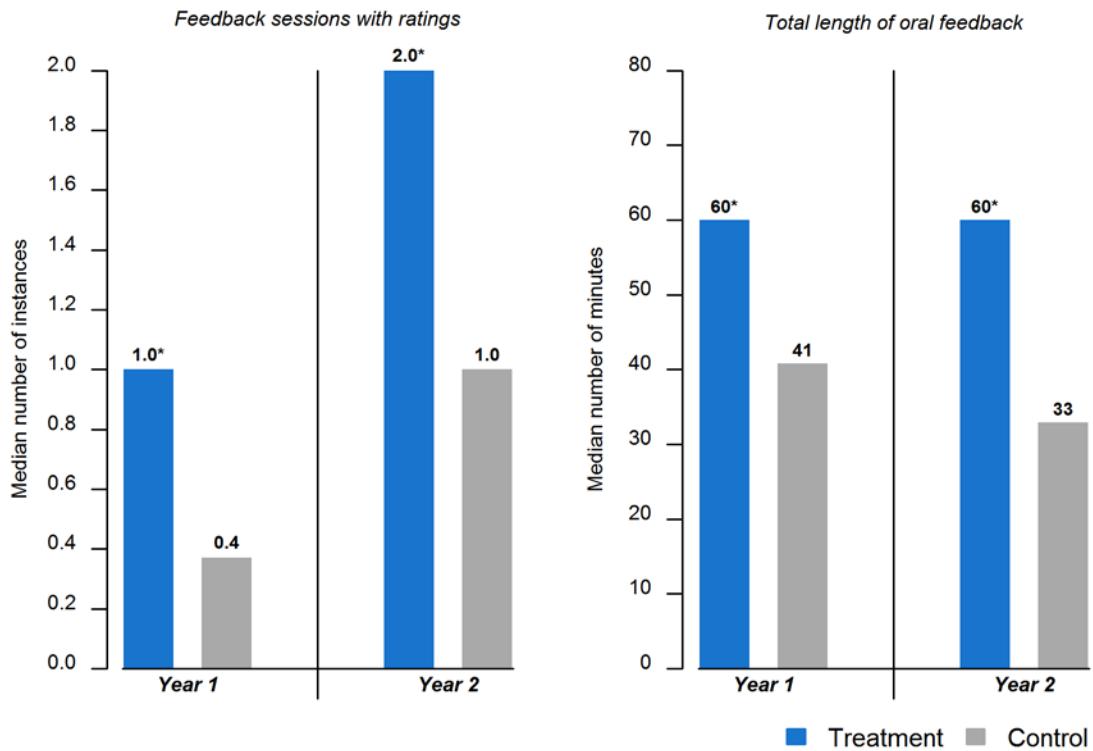


EXHIBIT READS: The average treatment principal in Year 1 reported receiving 1.0 feedback sessions with ratings, compared with 0.4 for control teachers.

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

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

SOURCES: Spring 2013 and Spring 2014 Principal Surveys.

Impact on Classroom Practice, Principal Leadership, and Student Achievement

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

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

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

Did the intervention have an impact on classroom practice?

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

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

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

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

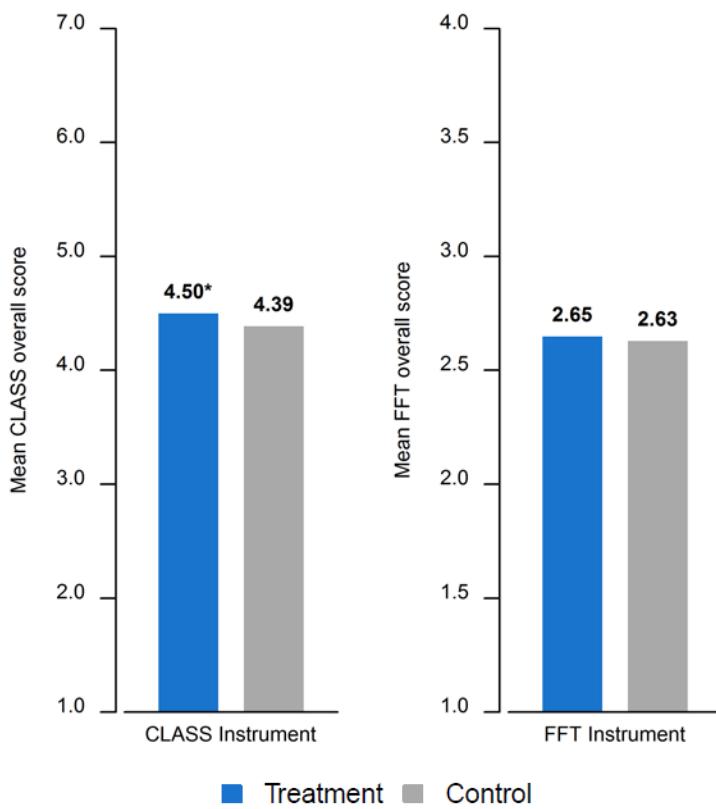


EXHIBIT READS: The average CLASS overall score was 4.50 for treatment teachers, compared with 4.39 for control teachers.

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

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

SOURCE: Spring 2014 Classroom Videos.

Did the intervention have an impact on principal leadership?

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

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

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

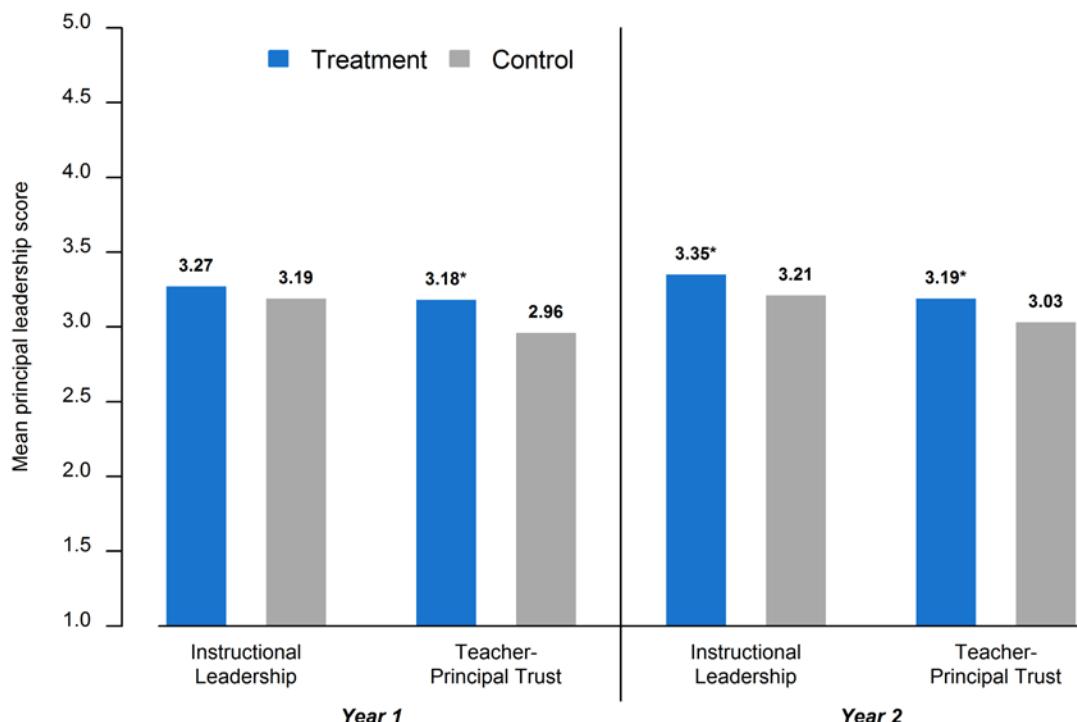


EXHIBIT READS: The average rating of principals' instructional leadership in treatment schools in Year 1 was 3.3, compared to 3.2 for principals in control schools.

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

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

SOURCES: Spring 2013 and 2014 Teacher Surveys.

Did the intervention have an impact on student achievement?

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

The intervention had a positive impact on students' mathematics achievement in Year 1, and had a cumulative impact similar in magnitude but not statistically significant ($p = 0.055$) in Year 2. The intervention did not have an impact on students' reading/English language arts achievement in either year. In Year 1, in mathematics, students in treatment schools scored at the 51.8th percentile in their district,

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

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

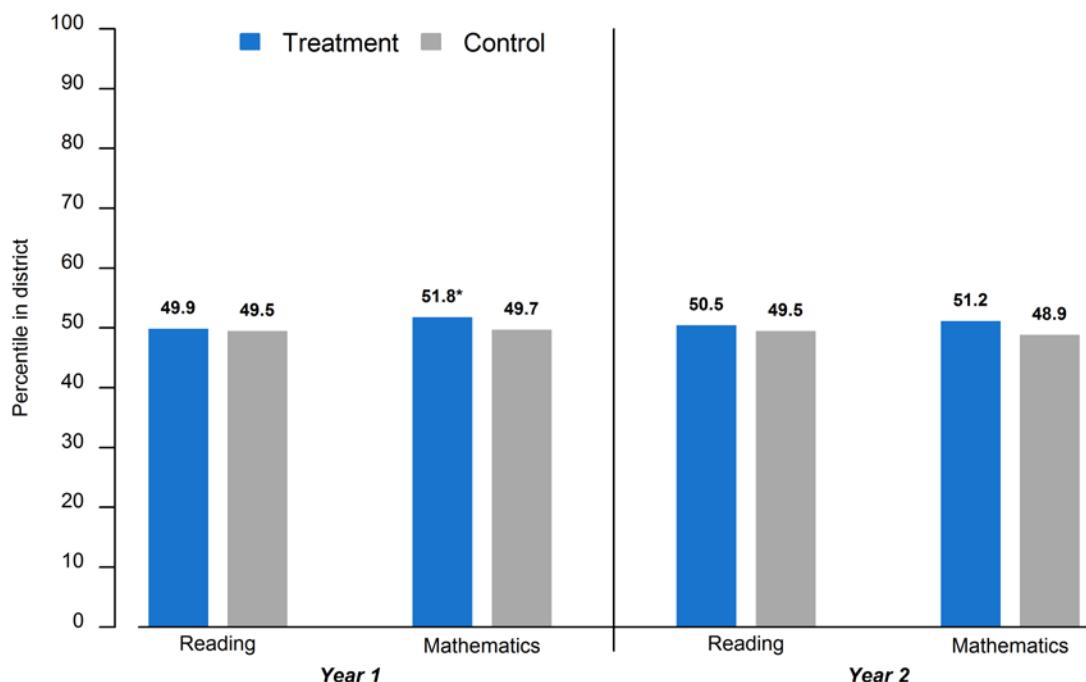


EXHIBIT READS: In Year 1, students in treatment schools received an average reading/English language arts score at the 49.9th percentile in their district, compared to the 49.5th percentile for students in control schools.

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

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

SOURCE: District Administrative Records.

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

Association Among Classroom Practice, Leadership, and Achievement

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

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

Chapter 1. Introduction

Educator performance evaluation systems are a potential tool for improving student achievement by increasing the effectiveness of the educator workforce.²⁴ For example, recent research suggests that providing more frequent, specific feedback on classroom practice may lead to improvements in teacher performance and student achievement.²⁵

This report is based on a study that the U.S. Department of Education's Institute of Education Sciences conducted on the implementation of teacher and principal performance measures highlighted in recent research, as well as the impact of providing feedback based on these measures.²⁶ As part of the study, eight districts were provided resources and support to implement the following three performance measures in a 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 goal is to examine the implementation of the intervention, including how fully it was implemented and the characteristics of the performance measures. These topics were the primary focus of the study's first report, which used data from the first year only.²⁷ The report shows that the educator performance measures were fully implemented, except many teachers and principals did not read the reports on teachers' contributions to student achievement growth. It also shows that the performance measures provided information with some but not all of the intended characteristics. For example, the ratings of classroom practice varied but were clustered in the top half of the scale, limiting their potential to signal a need for improvement. In addition, although the average of four classroom observations provided some information to identify teachers who needed support, individual observations had limited reliability to do so.

The study's second goal is to examine whether the intervention affected educator outcomes (e.g., teacher 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 addresses both goals, examining the impact of the two-year intervention, as well as implementation in both years. This chapter describes the intervention, research questions, and

²⁴ See Stecher et al. (2016); Weisburg et al. (2009).

²⁵ See Steinberg and Sartain (2015); Taylor and Tyler (2012).

²⁶ For recent research on performance measures, see, for example, Bill & Melinda Gates Foundation (2012, 2013).

²⁷ See Wayne et al. (2016).

design. Chapter 2 discusses the intervention’s three performance measures. It presents information about how fully they were implemented, the performance ratings each measure generated, and educators’ perceptions of the measures. Chapter 3 presents analyses of the impact of the intervention on the main outcome measures: teacher classroom practice, principal leadership, and student achievement. In addition, the chapter discusses whether the study design produced the intended contrast in performance feedback experiences and the intended impact on educators’ initial outcomes (e.g., educators’ perceptions of their own performance).

Overview of the Intervention

The intervention consisted of three types of 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., measures that vary across individuals and are reliable).
- Measures that provide clear and useful feedback at multiple times during each year.²⁸

In each of the eight participating districts, the intervention was implemented in a group of elementary and middle schools. A group of control schools in each district participated in the normal district evaluation processes only. To assist with implementation of the intervention, an American Institutes for Research (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 intervention specified how educators would receive the feedback (e.g., in feedback sessions after each observation). Other uses of the performance information were left to the discretion of the participating school and central office staffs. The study’s implementation team held meetings in each district to ask a group of school and central office educators to consider ways in which the performance information might be used—for example, to identify educators for praise or support, plan professional development, or guide coaching.

Districts were also given the option of using the information for staffing decisions (for example, decisions relating to tenure or continued employment). However, the study team anticipated that using the feedback for high-stakes purposes might be difficult, as it could 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

²⁸ Bill & Melinda Gates Foundation (2012, 2013); Whitehurst, Chingos, and Lindquist (2014).

observed that year. The study therefore tests the impact of providing feedback as an add-on to existing performance feedback, with no expected consequences (such as tenure or dismissal).²⁹

Below we describe each of the three intervention performance measures.

1. The Teacher Classroom Practice Measure and Feedback

This performance measure used classroom observations conducted four times during the course of each year, with a feedback session after each observation.³⁰ The intention was for one of the four observations to be conducted by an administrator from the teacher’s school, and for the other three to be conducted by study-hired observers (i.e., local professionals hired and trained by the study).³¹

After each observation, the observer was expected to prepare a report that included both ratings and narrative feedback on teacher 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.

Two different classroom observation systems were used to provide feedback. Districts were asked to choose between the Classroom Assessment and 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 the FFT.³² The use of two different observation systems was intended to make the study findings

²⁹ The available research evidence is mixed on whether stakes increase the effectiveness of feedback or attenuate it. Some researchers hypothesize, on the one hand, that employees may be more motivated to change their practices if they view their evaluation system as being used for the purpose of professional development rather than for dismissal (e.g., Atwater, Brett, and Charles 2007; Smither, London, and Reilly 2005). On the other hand, two recent studies in districts that provided feedback similar to that provided by this study’s intervention found that attaching stakes to the feedback had a positive effect. 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, the impact of attaching the prospect of a large financial bonus for sustained high performance. Using a regression discontinuity design, it found that both affected teachers’ performance ratings. These two studies were done in districts that provided feedback to all teachers similar to that provided by this study’s intervention and focused on the stakes attached to that feedback.

³⁰ In addition to four observations per year for the teachers who were the focus of the study (i.e., teachers of grades 4–8 who were responsible for reading/English language arts and mathematics instruction), the performance measure was used to provide two observations per year for teachers of kindergarten through grade 3—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 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 conducted, 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 grades K–3 teachers.

³¹ This distribution of effort was intended to engage principals in the implementation of the performance measure without overburdening them. Using multiple observers to rate the same teacher also produces a more reliable end-of-year average, compared with using a single observer for each teacher (see Ho and Kane 2013).

³² Several districts recruited for the study indicated that they had no particular preference for CLASS or FFT. These districts were assigned as needed to achieve the intended balance. We did not collect information on the reasons for the districts’ preferences.

more broadly relevant. 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.

The CLASS and FFT share many features that make them suitable for this 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. Second, there is evidence of validity and an association with student achievement for both instruments (Allen et al. 2011; Bill & Melinda Gates Foundation 2012; Goe, Bell, and Little 2008; Mashburn et al. 2010). Third, both instruments are applicable across subjects and grades. Finally, support for implementation was available from national vendors for both instruments. 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 and materials, web-based platforms for managing and reporting the performance information, and online video libraries with examples of teaching that exemplify particular levels of performance on each measured dimension.³³

2. The Student Growth Performance Measure and Feedback

This performance measure used student test results from multiple years to provide information about each teacher's contribution (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). It has been demonstrated that teacher value-added scores relate positively 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 et al. 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 student growth 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.³⁴ Computing value-added scores requires that students have at least one pretest score, so the student growth performance measure focused on grades 4–8 teachers who were responsible for instruction in reading/English language arts (ELA) and mathematics. All of the study districts had sufficient data to compute value-added scores in these grades.

An AIR team separate from the evaluation team designed and conducted the value-added analysis, drawing on AIR's experience doing similar work for states, as well as input from

³³ Two organizations provided support for the CLASS version of the classroom practice performance measure: Teachstone and the University of Virginia. Two organizations provided support for the FFT version of the classroom practice performance measure: Danielson Group and Teachscape.

³⁴ Treatment teachers were told during the study that they would be provided the third wave of value-added scores, based on the premise that the expectation that their contribution to student growth in the second year was being assessed might motivate improvement.

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 states and districts that measure value-added (see Collins and Amrein-Beardsley 2014). The model used for each district incorporated student test scores from the two prior years as predictors (where available), along with a set of measures of student characteristics selected by the districts. This choice of model and other design decisions was 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 E 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). The 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).³⁵ 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 expected to hold a one-on-one feedback session to discuss the results.

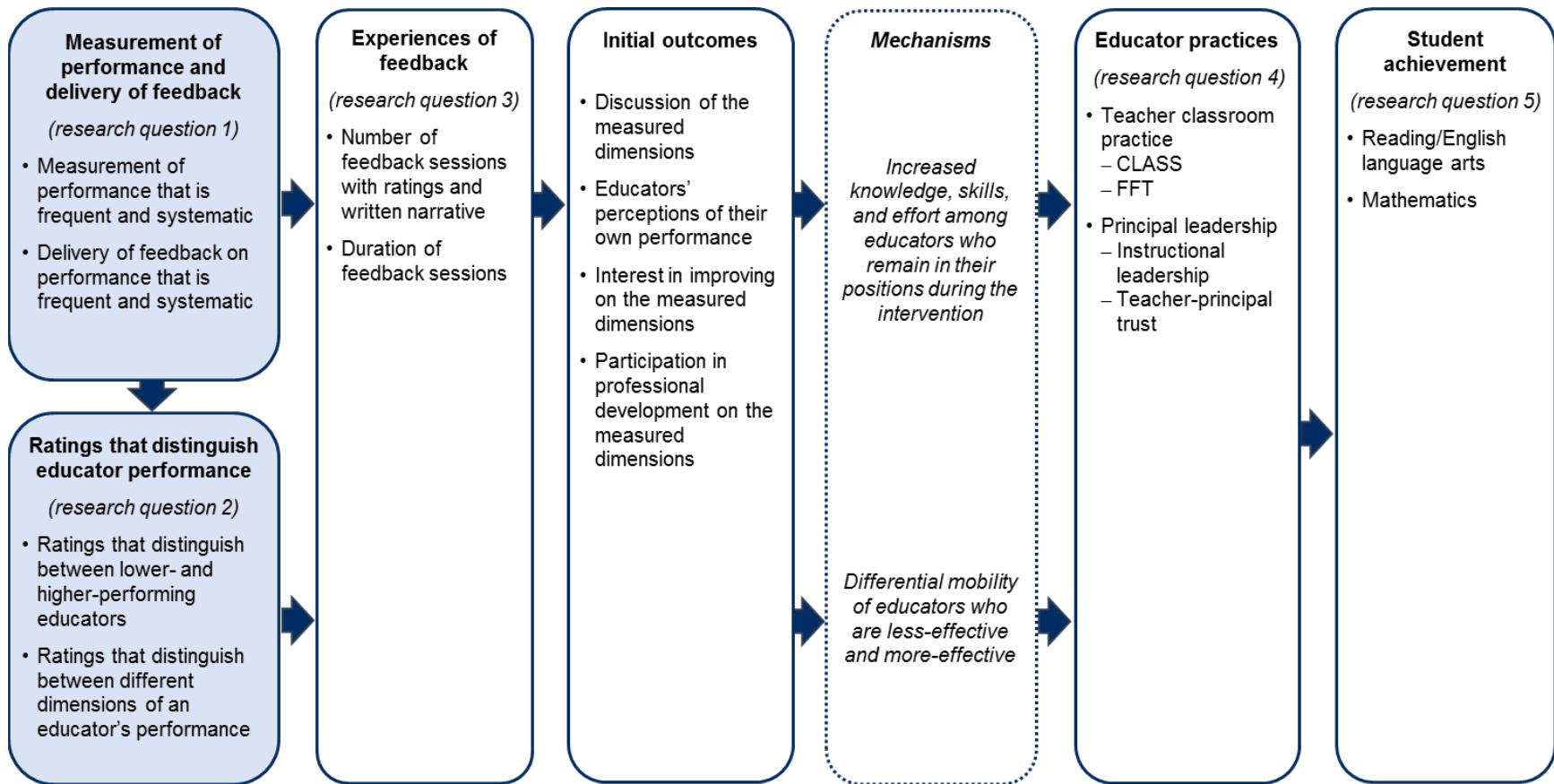
To prepare for implementation of all three performance measures, teachers, principals, and principal supervisors received trainings from the vendors. In addition, teachers received an orientation just prior to the beginning of the first study year. The orientation day included three hours on the intervention’s measures of classroom practice, one hour on the measure of student growth, and one hour on the measure of principal leadership. Just prior to the second study year, teachers and principals who were new to the study received the orientation, and continuing teachers received a half-day refresher on the intervention’s measures of classroom practice.

Theory of Action and Research Questions

This study is guided by a theory of action based on hypotheses about how performance measures and feedback affect the outcomes of educators—teachers and principals—and students. While there is some evidence that feedback on teachers’ performance can have an impact on student achievement (e.g., Steinberg and Sartain 2015; Taylor and Tyler 2012), there is little evidence on the intermediate mechanisms that lead to improved outcomes. The study’s theory of action begins with potentially important aspects of the implementation of the intervention (see shaded boxes on the left of exhibit 1.1) and continues with the experiences and outcomes that the intervention is expected to affect (see all other solid-line boxes on exhibit 1.1).

³⁵ The researchers who developed VAL-ED have published its psychometric properties in peer-reviewed journals and on their website (<http://www.valed.com/research.html>). See, for example, Porter et al. (2010).

Exhibit 1.1. Theory of action



According to the theory, *frequent and systematic performance measurement and feedback* may generate *ratings that distinguish* between lower- and higher-performing educators and between different dimensions of an individual educator's performance. This information could help identify educators in need of support, as well as the practices an educator should improve (see e.g., Donaldson and Papay 2014; Papay 2012). Providing this information to educators through feedback multiple times during the year could lead to ongoing improvement in their practices.

If educators *experience feedback* many times using the intervention's measures, the intervention may affect *initial outcomes*, including:

- *Discussions of the measured dimensions.* It may shift the focus of the feedback educators receive toward the measured aspects of classroom practice or leadership, causing increased discussion about those areas with supervisors and others who give feedback.
- *Educators' perceptions of their own performance.* It may lower some educators' perceptions of their effectiveness. The value-added scores provided by the intervention are expected to spread teachers across percentile ranks. That may lead some to think that they are less effective than they had thought. Research on teacher evaluation has noted that traditionally most teachers receive high ratings (Weisberg et al. 2009).
- *Interest in improving on the measured dimensions.* It may lead educators to want to become more effective in the measured areas of classroom practice or leadership because they perceive their performance as weaker than desired, because the feedback highlighted specific areas of practice as needing improvement, or because the feedback made them focus their attention on the measured practices.
- *Participation in professional development on the measured dimensions.* If they want to become more effective, they may seek out or be encouraged to participate in professional development on the measured dimensions.

In addition, the intervention may lead teachers to identify and try out new classroom practices independently or to reach out to colleagues informally for support.

If educators engage in these activities, it might affect teacher classroom practice and principal leadership through two *mechanisms*. First, it might cause *increased knowledge, skills, and effort* among teachers and principals who remain in their positions during the intervention. Second, positive feedback could lead higher-performing teachers and principals to remain in their schools, while negative feedback could lead lower-performing staff to leave, opening their positions to be filled by more-effective staff. Thus, the intervention could cause a *differential impact on mobility*, resulting in a more effective workforce.³⁶ Although the mechanisms provide an important link in the theory of action, the study design does not support inferences about the relative contribution of each mechanism.

Through those mechanisms, the intervention may have an impact on *educator practices*. According to the theory of action, the intervention may lead teachers to improve the specific classroom practices that are the focus of the intervention's classroom observation tool, as well as

³⁶ For literature discussing these mechanisms, see footnote 25 in chapter 1.

on other practices not as specifically targeted. In addition, the intervention may affect principal instructional leadership and teacher-principal trust, which are aspects of *principal leadership* that are associated with quality of instruction and student achievement (Sebastian and Allensworth 2012). By giving increased attention to teaching and learning (the focus of the VAL-ED performance measure) and by spending time observing and discussing classroom practices with teachers (the focus of the CLASS/FFT performance measures), the principal may become perceived by the teachers as a trusted instructional leader.

These improvements in classroom practice and principal leadership may lead to improved student achievement. The CLASS and FFT measures, like the leadership measure, have been shown to be related to improvements in student achievement (Allen et al. 2013; Kane and Staiger 2012). Thus, changes in classroom practice and principal leadership may lead to improved *student achievement*, as shown in the far right of the theory-of-action diagram. (See exhibit 1.1.)

This multiyear study is designed to examine the implementation of an intervention that is guided by this theory of action, and to 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 identify more and less effective educators and signal the specific dimensions of practice that most needed improvement?
3. To what extent did educators' experiences with performance feedback differ for treatment and control schools?
4. Did the intervention have an impact on teacher classroom practice and principal leadership?
5. Did the intervention have an impact on student achievement?

This report will address all five questions, spanning both study years.

Overview of Study Design

To answer the research questions, we recruited a sample of eight districts and conducted the study in a 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. Both the treatment and control groups continued to implement the districts' existing educator evaluation systems. In participating schools, the study focused on the principals and teachers of reading/ELA and mathematics in grades 4–8.³⁷

³⁷ Teachers of kindergarten through grade 3 also participated in the study. This was done mainly to promote schoolwide engagement in the implementation of the teacher 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.

This section describes how we selected suitable districts and schools, how we randomly assigned schools to treatment and control groups, the data we collected, and the analytic methods we used.

Districts and Schools

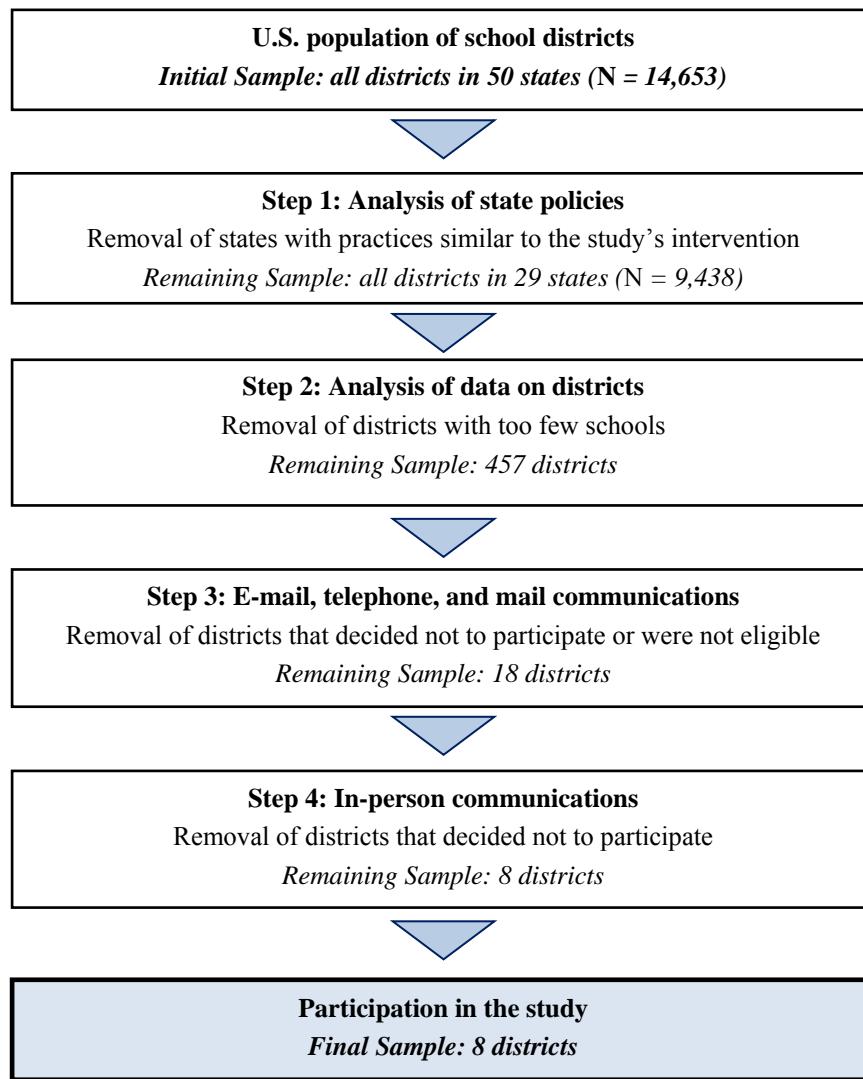
The study was conducted in a sample of districts and schools. This sample was selected purposefully, based on criteria established by the study team. This subsection describes how we selected districts, the districts' characteristics, and the districts' performance feedback practices for their existing educator evaluation systems. It also describes how we selected schools, as well as the schools' characteristics.

District 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 with an analysis of state policies for the evaluation of teachers and principals. (See exhibit 1.2.) 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 (fall 2012 to spring 2014). The study team excluded districts from those states. Although many other states intended for their districts to implement such practices because of the Elementary and Secondary Education Act Flexibility Waivers, full implementation was not required until fall 2014 at the earliest. For this reason, districts in many states were eligible for the study despite the state's participation in the waiver program.

Within the 29 states that were eligible to participate, 457 districts 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, 49 of which expressed interest in a follow-up conversation about participating. The study team assessed district eligibility and determined that some were not eligible, either because they did not have data systems that made the student growth performance measure feasible or because they had policies for evaluating teachers and principals that did not contrast with the intervention's performance measures. Of the 36 districts 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 teachers' organization or the aggressive schedule to begin implementation in summer 2012).

Exhibit 1.2. District selection and recruitment process



District Characteristics. 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. This was primarily due to the removal of districts that did not have the required number of schools to participate.

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: Percentages for characteristics with multiple categories may not sum to 100 because of rounding.

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

The sample also included districts with different state policies with respect to collective bargaining. Three of the eight districts (37.5 percent) were in states where collective bargaining is required. (To provide a point of comparison, 67.7 percent of districts across the United States are in states where collective bargaining is required.) Two were in states where collective bargaining is permissible, and three were in states where it is illegal. During the final step of the recruitment process, some districts in states requiring collective bargaining decided not to participate. Although it is not possible to know districts' reasons for choosing not to participate, it was common for districts with collective bargaining agreements to consider teacher union support as a factor in their decision making.

Performance Feedback Typically Provided in the Districts. By design, the performance feedback provided as part of the intervention was to be given in addition to the feedback typically provided by districts. We conducted interviews with each district to determine what type of feedback they typically provide. (The interviews are described further in the section titled “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 intervention.

Districts’ feedback on classroom practice. All eight study districts required less frequent observations 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 terms of who conducted the observations. Under the districts’ evaluation systems, school administrators conducted the observations. In contrast, the intervention used study-hired observers for three of the four observations each year. District policies also differed from the intervention in terms of the training requirements for observers. The districts required an average of 13.5 hours of training—a little over half of the duration of the study’s training.³⁸ In two of the eight districts, no observer training was required. Only three districts required observers to pass an assessment of rating skill, which was 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. (For instance, the names of the performance levels may have been changed or the text that defines the performance levels for each dimension may have been altered.)

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.)

³⁸ The required observer training for the study’s intervention was 20 hours for observers in the CLASS districts and 26 hours in the FFT districts.

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

District ID and assigned classroom observation system for intervention		Districts' feedback on teacher classroom practice						Districts' feedback on student growth	
		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					

NOTES: ^aNumber of observations shown is the minimum required under each district's evaluation system. Administrators could observe more frequently at their discretion.

^bEach 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.

^cThe 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 Interviews.

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 intervention's plan of twice a year. (See exhibit 1.5.) District policies for principal evaluation also differed from the intervention in terms of the nature of the information used for feedback: None of the districts used the VAL-ED instrument (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. 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

District ID and assigned classroom observation system for intervention		Districts' feedback on principal leadership			
	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

NOTE: ^aData for two districts are missing because the districts did not provide the rating instruments.

SOURCE: District Interviews.

School Selection and Characteristics. Each of the eight districts identified a set of schools that met the study's eligibility criteria and agreed to participate. The study's focus on teachers of reading/ELA and mathematics in grades 4–8 meant that only elementary and middle schools were eligible to participate. To reduce heterogeneity, the school sample was also 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. 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 appendix exhibits A.1 and 2; for the characteristics of schools in the districts that used CLASS and FFT, see appendix exhibit A.3.)

Random Assignment of the Schools

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

To maximize the precision with which the study could compare outcomes in the treatment and control groups, random assignment was conducted separately within 37 blocks. The blocks were defined by district and school level (elementary schools or middle schools), so that half of each district's elementary schools were treatment schools and half were control schools, and half of each district's middle schools were treatment schools and half were control schools. Blocks also 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 we examined: the percentage of principals with 4–10 years of experience. This percentage was lower for treatment principals than for control principals by a statistically significant amount (17 versus 33 percent). (See appendix exhibits A.4a–j.)³⁹

One control school from the first study year did not continue to participate in the second year because the school was restructured.

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

³⁹ Appendix A also includes baseline equivalence results for the CLASS districts and the FFT districts separately.

Data Collection

The study collected the following data on the implementation of the intervention and the information provided to teachers and principals in the treatment schools:

Implementation of the measures. We documented attendance at orientation and training events related to the study's performance measures. Online system records maintained by the vendors provided 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 implementation of the observations and the district context, respectively. Finally, surveys of teachers and principals administered in the spring of Year 2 asked about perceptions of the performance information received from the study's classroom observation and principal leadership practices measures compared to that received from the districts' official performance system. Both groups also reported on their perceptions of the information from the student growth measure (e.g., whether it was easy to understand and a good measure of how well students had learned).

Information provided to teachers and principals. The data generated by the measures of teacher classroom practice, student growth, and principal leadership were collected through the vendors' online systems.

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

Educators' experiences with performance feedback. In the spring of each study year, we surveyed the teachers and principals in treatment and control schools to collect information on the nature and frequency of the performance information educators received, as well as their perceptions of that information.

Initial outcomes. The spring surveys also asked about initial outcomes, including whether teachers and principals wished to improve or sought professional development in areas covered by the feedback. The surveys also asked teachers and principals for perceptions of their own performance.

Finally, we collected data on three main outcomes:

Teacher classroom practice. To provide a common measure of classroom practice in treatment and control schools, we video-recorded each teacher's instruction in the spring of Year 2. We video-recorded one lesson per teacher and then selected a random sample of half of the respondents for a second round of recording.⁴⁰ We coded each of the videos using the CLASS

⁴⁰ We video-recorded two lessons for some teachers and one for others to achieve the desired precision while minimizing cost and burden.

and FFT.⁴¹ This allowed us to examine the impact on a measure of practice aligned with the measure selected for feedback and on a measure that was similar, but not completely aligned.

Principal leadership. To provide a common measure of principal leadership in treatment and control schools, we relied on teachers' responses to survey items designed to assess principals' instructional leadership and teacher-principal trust, based on measures developed by the Chicago Consortium on School Research (CCSR, 2012).⁴²

Student achievement. To measure student achievement, we collected students' scores on state standardized tests in reading/ELA and mathematics.

In addition to the collections described above, we collected data on the characteristics of principals, teachers, and students in study schools from district administrative records in the summer and fall of 2012, fall 2013, and fall 2014.

Response rates for the data collections were high. The response rate for the videotapes of classroom practice was 91.6 percent. Every other data collection achieved a response rate of nearly 100 percent. In the second study year, for example, the overall response rate was 98.6 percent for the teacher survey and 96.0 percent for the principal survey. (Additional details on data collection and response rates appear in appendix B.)

Analytic Approaches

This section discusses the analytical methods used to examine implementation and outcomes. We refer to the first study year (2012–13) as Year 1 and the second year (2013–14) as Year 2.

Implementation of the Intervention. To examine implementation of the intervention, we conducted descriptive analyses of the extent to which study participants in the treatment group received the intended training on the performance measures, carried out the anticipated 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., percentage of principals with an overall rating of *distinguished*) and the correlations among different performance measures. In addition, we used a generalizability theory framework (Shavelson and Webb 1991) to estimate the reliability of the performance scores educators received. Within this framework, reliability is defined as the proportion of variation in a measure's scores that reflect "true" differences

⁴¹ To the extent possible, video-recording was scheduled to take place when a teacher was teaching either reading/ELA or mathematics. Overall, 45 percent of the video-recorded lessons were in reading/ELA, 50 percent in mathematics, and 5 percent in other subjects.

⁴² It was not feasible to use the VAL-ED itself as an outcome measure. By the time of the Year 2 spring surveys, a large majority of treatment teachers had already completed the VAL-ED four times, making it likely that they would respond to the survey with a disposition or framework different from that used by control teachers, who had never before completed a VAL-ED survey.

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

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

Impact of the Intervention. For analyses of the impact of the intervention in Year 1 and Year 2, we focused on the principals, teachers, and students present near the end of each school year (i.e., the “impact sample”). Any statistically significant differences in values between the treatment and control participants in the impact sample can be interpreted as impacts.

As expected, some members of the impact sample joined during the two-year period of implementation, replacing principals and teachers who had left. Among those present in the Year 2 impact sample, 17 percent of principals and 25 percent of teachers were not present in the Year 1 impact sample.⁴³ These movements do not affect the internal validity of the study’s inferences about the impact of the intervention because the movements are one mechanism through which the intervention may have an impact, as shown in the theory of action. (See exhibit 1.1; for detailed charts showing principal, teacher, and student movements during the study, see appendix exhibits A.5–8.)

Based on the impact samples, we assessed the impacts of the study’s intervention on different types of outcomes using different statistical models, as summarized below.

- To assess the impact on educators’ experiences with performance feedback, we compared the means for the treatment and control groups using a two-level linear probability model for binary measures (e.g., whether a teacher received feedback based on observations). For continuous measures of educators’ experiences (e.g., the number of instances of feedback received), we compared the median rather than the mean for the treatment and control groups. We did so because many of the survey-based continuous measures were not normally distributed.⁴⁴

⁴³ These rates did differ by treatment condition. The percentages of treatment and control principals in the Year 2 impact sample who were not present in the Year 1 impact sample were 21 percent and 14 percent, respectively. For teachers, the percentage in the Year 2 impact sample who were not present in the Year 1 impact sample were 22 percent and 28 percent, respectively. (For further details, see appendix exhibits A.5 and 6.)

⁴⁴ 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.

- To assess the impact on teachers' initial outcomes, we used survey data (e.g., their self-ratings and their interest in improving specific areas of practice) to estimate a two-level linear model (with teachers nested within schools).
- To assess the impact on teacher classroom practice, we used observation data to estimate a three-level model (with lessons nested within teachers—one or two lessons per teacher depending on the number of lessons sampled—and teachers nested within schools).
- To assess the impact on principals' initial outcomes, we conducted a principal-level linear regression using principal survey data (e.g., their self-ratings and their interest in improving specific areas of practice).
- To assess the impact on principal leadership, we used teacher survey data to estimate a two-level model (with teachers nested within schools).
- To assess the impact on student achievement, we used a three-level model (where students are nested within teachers and teachers nested within schools) with data pooled across grades 4–8.

For all impact analyses, the models accounted for random assignment blocks and, where applicable, the nesting of students within teachers and teachers within schools. In addition, analyses of impacts on educators' initial outcomes, teacher classroom practice, principal leadership, and student achievement incorporated a set of covariates (e.g., student and teacher background characteristics) to improve the precision of the impact estimates and adjust for any baseline differences between the study groups. Detailed descriptions of each model are provided in appendix H. Appendix H also includes descriptions of additional analyses that we conducted to determine the sensitivity of the main impact results to alternative model specifications.

In addition to the analyses described above, we checked whether the impact results differed across subgroups of principals and teachers. Specifically, we tested whether the effects differed for probationary and nonprobationary teachers, teachers in elementary and middle schools, and teachers with lower and higher value-added scores. (See appendix H for details.)

Finally, to supplement the impact analyses, we examined the association between classroom practice and principal leadership with student achievement. These relationships were estimated by adding each measure of classroom practice or principal leadership as a predictor to the main student achievement impact model. (See appendix H for details.)

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

Chapter 2. Implementation of the Performance Measures and Feedback

This chapter discusses the design and implementation of the intervention’s three performance measures. For each, it describes the measure’s design, how fully it was implemented, and how well it differentiated educator performance, all of which may affect the usefulness of the measure. The chapter also examines teachers’ and principals’ perceptions of the feedback they received, including whether they reported that it provided clear ideas about how to improve. All findings in this chapter pertain to teachers and principals in the treatment schools only.

Key Findings About Implementation

Measures of Classroom Practice

- Teachers received nearly all the intended feedback sessions each year.
- Nearly all teachers had overall classroom observation scores in the top two performance levels, limiting the potential of the information to signal a need for teachers to improve.
- Teachers’ overall classroom observation scores—averaged across all four windows in a year—provided some reliable information for identifying teachers who needed support, but single observations did not. In addition, the observations did not reliably indicate areas for improvement.
- Three-quarters of treatment teachers said that the study’s feedback on classroom practice was better than previous feedback from the district, averaging over six characteristics (e.g., more useful, more specific).

Measure of Student Growth

- In the first year, less than half of the treatment teachers and principals viewed their student growth reports, which were available through a secure web portal. In the second year, hard copies of the reports were disseminated, and almost all treatment teachers and principals received their reports.
- For just under a quarter of teachers with value-added scores in reading/ELA, and about half with value-added scores in mathematics, the scores measurably differed from the district average, thus providing some reliable information to signal whether a teacher needed to improve.
- Only about half of the teachers reported positive perceptions of the reports they received.

Measure of Principal Leadership

- Nearly all treatment principals received two VAL-ED feedback sessions each year.
- Principals were spread across the full range of performance levels, consistent with the VAL-ED norms.
- VAL-ED ratings provided by principals, supervisors, and teachers in the two fall administrations were often too different to form a reliable measure, but the spring ratings were consistent enough to indicate whether a principal needed to improve.
- On four of five items, nearly three-quarters of principals said the study’s feedback on leadership was better than previous feedback from the district (i.e., easier to understand, more objective, more specific about what high quality is, and provided clearer ideas about improving leadership); however, over half of the principals (55 percent) reported that the study’s feedback was less comprehensive.

Supplemental tables for the chapter appear in appendixes D, F, and G, which each focus on one of the intervention’s three performance measures.⁴⁵ Samples of the reports on teachers’ and principals’ performance appear in appendix K. This chapter is based on analyses of implementation in both study years. It builds on the findings outlined in the first report (Wayne et al. 2016), which explored implementation in Year 1 in detail. The implementation findings were similar across years, with a few exceptions. The chapter shows all results for both Year 1 and Year 2, except where noted.

The Intervention’s Measures of Teacher Classroom Practice

Overview of the Measures

Districts were given the opportunity to choose between two rating systems for measuring classroom practice, as described in chapter 1. Four districts chose CLASS and four chose FFT. In this section, we present implementation results for the eight districts together, as well as for the CLASS and FFT districts separately.⁴⁶ The study did not randomly assign districts to use CLASS or FFT, which means that differences in results between the CLASS and FFT districts cannot necessarily be attributed to the observation systems; such differences could occur due to other district characteristics.

The CLASS and FFT versions of the intervention’s teacher classroom practice measures were designed to provide information on multiple dimensions repeatedly throughout each year. Specifically, they were designed to include the following features:

- Four observations during each school year, one conducted by the principal or another school administrator, and three conducted by study-hired observers, scheduled such that teachers knew the week when they would be observed, but not the day or time.^{47,48}

⁴⁵ Appendixes D, F, and G contain several additional results, for reference. Appendix D contains all results disaggregated according to whether the district used the CLASS or FFT version of the study’s feedback on classroom practice, when not shown in the chapter. Analyses of the implementation of the measures of classroom practice in the report are based on teachers of grades 4–8, which were the main focus of the intervention. Results for teachers in grades K–3 corresponding to exhibits 2.2 and 2.3 appear in appendix D. Teachers of kindergarten through grade 3 in treatment schools participated in some aspects of the intervention to promote schoolwide engagement (see chapter 1). 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.

⁴⁶ Findings on the implementation of the feedback on student growth and principal leadership are presented for CLASS and FFT districts separately in appendixes F and G.

⁴⁷ 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 4-window average scores. However, we concluded that the potential benefits of rapport would outweigh the improved reliability.

⁴⁸ To the extent possible given the constraints of scheduling, the principal and study-hired observers were asked to conduct the four observations for each teacher when the teacher was teaching the same subject (either reading/ELA or mathematics) and during the same class period. Conducting observations during the same subject and class period was intended to make it easier for teachers and principals to interpret the observation ratings. In addition, within

- A report prepared by the observer after each observation, including ratings and narrative feedback.
- An in-person feedback session after each observation, during which the observer reviews the report with the teacher.

The two systems capture similar dimensions of classroom practice and involve similar feedback sessions. However, they differ in terms of the amount and kind of information on teacher performance provided in the reports.

The CLASS districts used the upper-elementary version of CLASS, which covers 12 dimensions of classroom practice grouped into four domains. (See exhibit 2.1.)⁴⁹ All scores are on a 7-point scale. The FFT is designed for use in grades K–12. The FFT has four domains, two of which can be scored based on classroom observations. The FFT districts used only those two domains, which together include 10 dimensions of classroom practice. (See exhibit 2.1.) All scores are on a 4-point scale.

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

Classroom Assessment and Scoring System (CLASS—Upper Elementary)	Framework for Teaching (FFT)^a
Domain 1: Emotional Support <ul style="list-style-type: none"> • Positive climate • Teacher sensitivity • Regard for student perspectives 	Domain 2: Classroom Environment <ul style="list-style-type: none"> • Creating an environment of respect and rapport • Establishing a culture for learning • Managing classroom procedures • Managing student behavior • Organizing physical space
Domain 2: Classroom Organization <ul style="list-style-type: none"> • Behavior management • Productivity • Negative climate 	Domain 3: Instruction <ul style="list-style-type: none"> • Communicating with students • Using questioning and discussion techniques • Engaging students in learning • Using assessment in instruction • Demonstrating flexibility and responsiveness
Domain 3: Instructional Support <ul style="list-style-type: none"> • Content development • Quality of feedback • Analysis and inquiry • Instructional dialogue • Instructional learning formats 	
Domain 4: Student Engagement <ul style="list-style-type: none"> • Student engagement 	

NOTES: ^aThe 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.

During the feedback sessions, the observers were expected to focus on two or three dimensions, including at least one strong and one weak dimension. For each dimension, the observers were expected to talk about the behavioral indicators associated with the teacher’s score, as well as

each school, the study-hired observers were encouraged to balance the number of teachers who were observed during reading/ELA and mathematics, if feasible.

⁴⁹ 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.

those associated with a higher score. Observers would then discuss actions the teacher could take to earn a higher score.

The CLASS and FFT online platforms were designed to provide each teacher with a report for each observation, which the observer would review with the teacher during the in-person feedback session.⁵⁰ The reports generated by the online platforms differed in content. CLASS reports provided separate scores for individual dimensions, as well as the teacher's overall score and a sense of how his or her performance compared with others. The FFT reports only provided separate scores for individual dimensions. (For sample CLASS and FFT reports, see appendix K.)

Implementation of the Measures of Classroom Practice

The implementation team worked with each district to identify and hire observers to conduct the observations and feedback sessions consistent with the study design. Observers received the standard training offered by the CLASS and FFT vendors to learn how to reliably score instruction, enter scores and narrative text for the reports, and conduct feedback sessions with teachers. All observers passed the certification test, demonstrating that they could score reliably.⁵¹ In spring of Year 2, almost all principals reported that they felt prepared to rate instruction and provide feedback using the study's measure of classroom practice. For example, 100 percent reported that they had a clear idea of what the study's rating system for classroom practice defines as good instruction.⁵² (For additional results, see appendix exhibit D.1.)

Each teacher was to be observed and provided feedback throughout the year, once during each of the four calendar windows defined by each district. Although the goal was four rounds of observation and feedback per year, a teacher who was observed in the first windows of the school year could leave in the winter and be replaced. In that scenario, the replacement teacher would receive feedback only for the remaining windows for the year. A replacement teacher who joined in the summer between Year 1 and Year 2 would receive four observations at most, all occurring during Year 2.

Teachers received nearly all the intended feedback sessions each year. The average number of feedback sessions received each year by teachers present in the spring was 3.7 sessions in Year 1 and 3.9 sessions in Year 2. This means that teachers received close to the intended dosage of four feedback sessions each year. Teacher mobility and other implementation challenges did not lessen the dosage by much in either year. (See exhibit 2.2.)

Each feedback session may spur additional improvements in classroom practice. For this reason, it is also important to assess the *cumulative dosage* received by those present at the end of Year 2. The cumulative dosage was close to what was intended. The teachers present in spring of

⁵⁰ The CLASS and FFT online platforms were also equipped to provide each principal with reports on all of the teachers he or she supervises.

⁵¹ See chapter 2 of Wayne et al. (2016) for more details on the characteristics of observers, their training, and experiences with the certification test.

⁵² This finding is based on survey items that appeared in the Year 2 survey only, so the appendix does not contain parallel results for Year 1.

Year 2 received an average of 6.8 feedback sessions, instead of the intended eight sessions. These results are close to what would be expected, given the mobility rates of treatment teachers: 23 percent of teachers present in spring of Year 2 were not present at the beginning of Year 1. Almost all of these teachers transitioned in during the summer between the two study years and therefore received four observations in the second year.⁵³

Exhibit 2.2. Mean number of feedback sessions treatment teachers received in each study year and in total

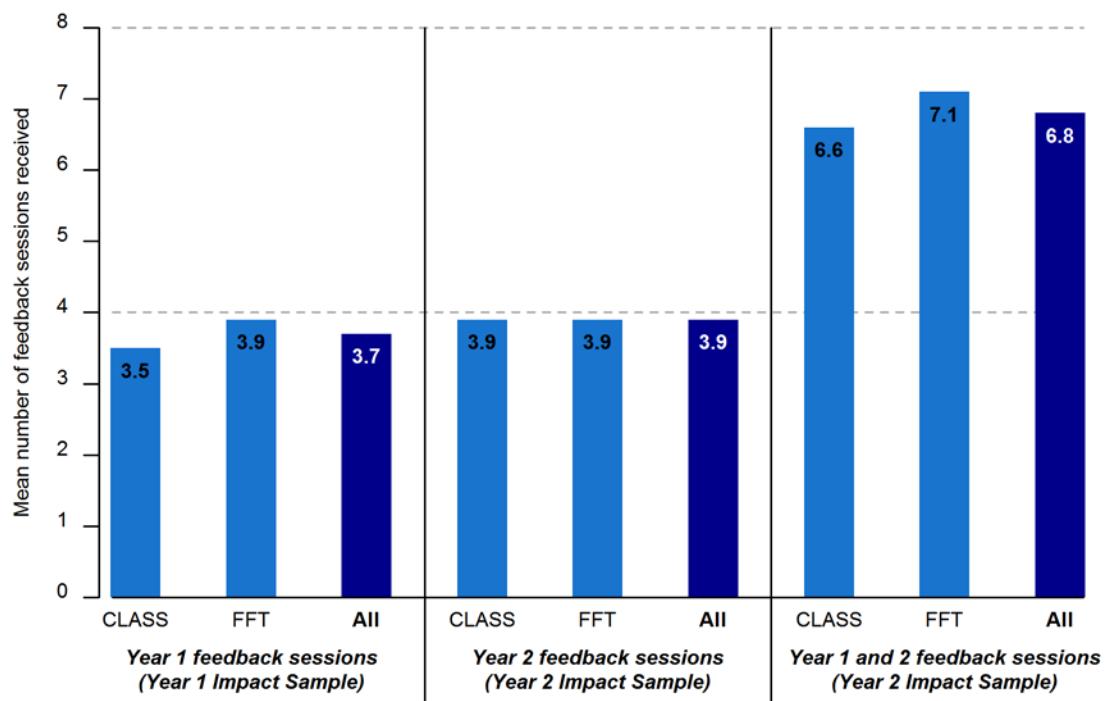


EXHIBIT READS: On average, treatment teachers in the Year 1 impact sample in CLASS districts received 3.5 feedback sessions in Year 1.

NOTES: Sample size for Year 1 = 527 teachers (308 CLASS and 219 FFT). Sample size for Year 2 = 504 teachers (305 CLASS and 199 FFT). See appendix exhibit D.2 for results for grade K–3 teachers.

SOURCES: Teachstone Online System and Teachscape Online System.

The feedback sessions were supposed to engage teachers and help them understand the feedback, identify practices to improve, and think about what improved practice would look like. During the sessions, observers could show video clips—drawn from the CLASS or FFT provider’s online video library—to illustrate strong performance on a specific dimension of practice. CLASS observers were expected to show the teacher one or two clips relevant to the dimensions that were a focus in the feedback and recommend additional videos for the teacher to view on his or her own. FFT observers were not told to show videos in the feedback sessions; instead, they were told to recommend videos and other resources on the Teachscape website. Teachers could then review these after the feedback session to help them think about how to improve their

⁵³ A total of 3 percent of the teachers present in the spring of Year 1 were not present at the beginning of Year 1; likewise, 5 percent of teachers present in spring Year 2 were not present at the beginning of Year 2.

instruction. Recognizing that viewing video clips could increase teacher engagement in the feedback sessions and make the sessions more useful, we asked study-hired observers if they *typically* showed video clips, where *typically* is defined as using the tools in two-thirds or more of the feedback sessions they conducted.⁵⁴ In both years, about 60 percent of study-hired observers in CLASS districts and less than 10 percent in FFT districts reported that they typically showed video clips in their feedback sessions. (For detailed results, see appendix exhibit D.3.) We did not obtain information from teachers about the use of video clips in feedback sessions.

A large majority of study-hired observers reported that teachers typically appeared engaged and interested during the feedback sessions. A large majority of study-hired observers in both groups of districts reported that it was typical for teachers to be actively engaged in discussions during feedback sessions: across all districts, 80 percent reported this in Year 1 and 92 percent did so in Year 2. (For detailed results, see appendix exhibit D.4.) The teacher survey did not include items about teachers' engagement level during the feedback sessions.

Performance Information on Classroom Practice

As described earlier in the chapter, the classroom practice measure included detailed information for teachers about their teaching. 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 10 FFT dimension scores, each of which was on a 1–4 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*).

In this subsection, we begin by examining whether the overall scores identified teachers as needing support: We first describe the variation in overall scores within each observation window and in average scores across the four observation windows;⁵⁵ we then examine whether the classroom practice scores distinguished among teachers whose persistent performance during the year was better or worse. The subsection also presents findings on how well the teachers' reports identified the dimensions of practice they most needed to improve.

Nearly all teachers had overall classroom observation scores in the top two performance levels, limiting the potential of the information to signal a need for teachers to improve. For CLASS observations, nearly all of the teachers (98 percent or more) received an overall score that placed them in the top two performance levels within each observation window in Year 2, labeling them *effective* or *highly effective*. The distributions of

⁵⁴ The study-hired observer surveys, administered in the spring of each year, included items on how frequently the observer used these tools. The survey items reported here focused on the feedback sessions conducted between January 1 and the survey completion date. For a given approach, the study-hired observers could respond "One or two," "Some (more than two, up to one-third)," "Many (between one-third and two-thirds)," "Nearly all (more than two-thirds)," or "All."

⁵⁵ The "4-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. For teachers who had fewer than four observations, the average score is based on the number of observations they had during the year.

teachers across performance levels for windows 1 and 4 appear alongside the distribution for the four-window average score. (See exhibit 2.3.) For FFT (depicted on the right side of the exhibit), more than 87 percent of the teachers within an observation window had an overall score of 2.50 or higher, which corresponds to the top two study-defined performance levels. These distributions are very similar to the distributions for Year 1.

Exhibit 2.3. Distribution of teachers across CLASS and FFT performance levels for Windows 1 and 4 and for the 4-Window average, Year 2

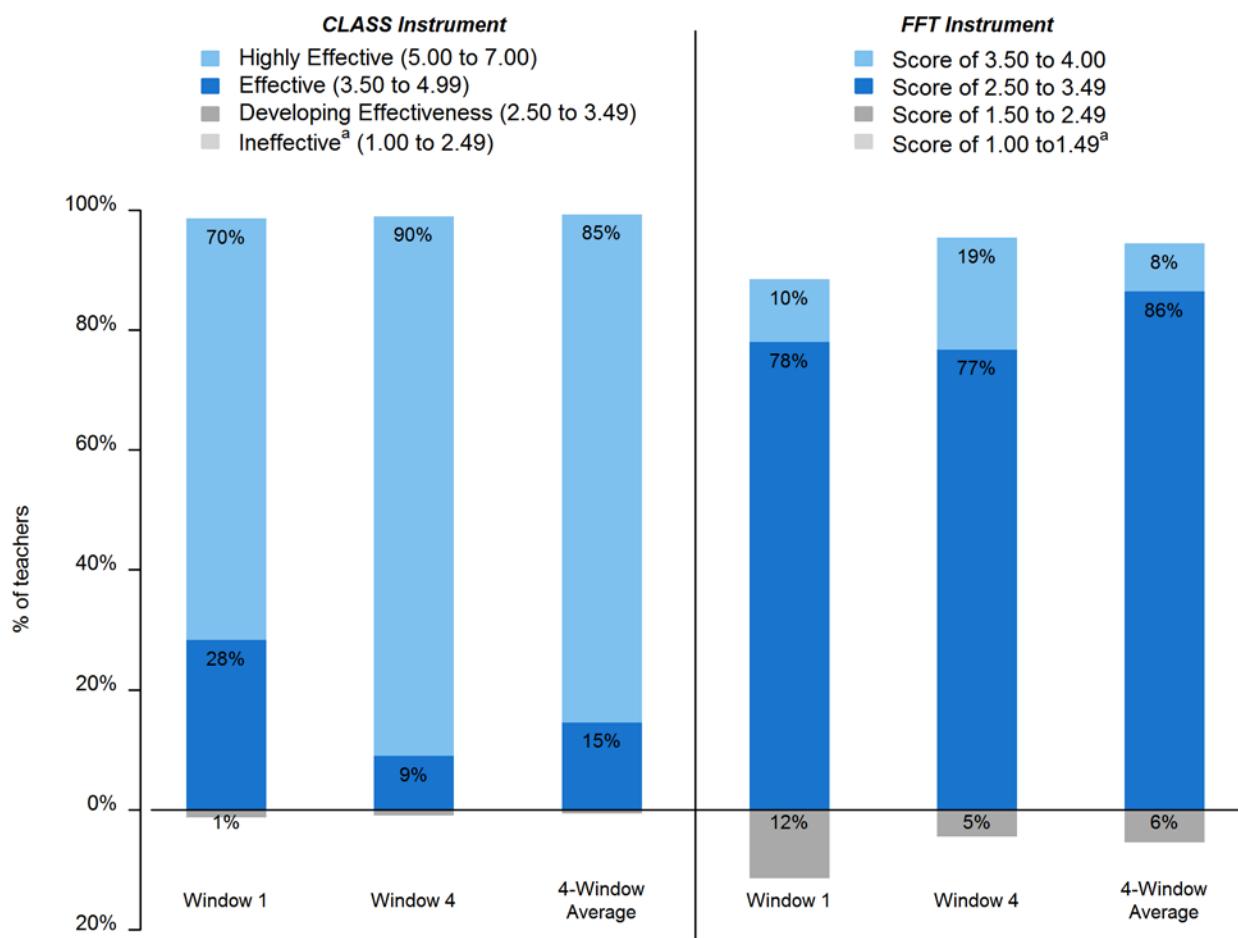


EXHIBIT READS: Of treatment teachers in CLASS districts observed in Year 2 window 1, 70 percent had a CLASS overall score at the *highly effective* performance level, 28 percent at the *effective* performance level, and 1 percent at the *developing effectiveness* performance level. Less than 1 percent of teachers had an overall score at the *ineffective* performance level.

NOTES: Sample size for CLASS districts = 297 teachers in Window 1, 302 teachers in Window 4, 303 teachers for the 4-Window average. Sample size for FFT districts = 191 teachers in Window 1, 198 teachers in Window 4, 199 teachers for the 4-Window average. Percentages for each window and for the 4-Window average may not sum to 100 percent due to rounding. See appendix exhibits D.6a and 6b for results for grade K-3 teachers.

^aWithin a window, in the CLASS as well as the FFT districts, less than 1 percent of teachers had an overall score at the lowest of the four possible score bands (i.e., the CLASS band from 1.00 to 2.49, and the FFT band from 1.00 to 1.49).

SOURCES: Teachstone Online System and Teachscape Online System

Although overall classroom observation scores were concentrated toward the high end of the rating scale, they still varied across teachers. Even among teachers with the same performance-level designation, the overall score distributions indicate that there were differences in teachers'

overall scores.⁵⁶ In addition, scores rose between the first and fourth window each year, but fell between years, such that on average a teacher's scores for the fourth window of the first and second year were similar.^{57,58}

The overall score averaged across four windows provided some reliable information to identify teachers who needed support. However, 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 measure average performance over the course of each year *reliably*. If reliable, a teacher's overall scores reflect persistent classroom practice rather than idiosyncratic factors introduced by the observer or the particular days or lessons observed.⁵⁹ Educators should have more confidence in decisions and actions based on more reliable measures, although what constitutes "sufficient" reliability depends on the measure's intended use.⁶⁰

We estimated the degree to which a teacher's 4-window average score was a reliable measure of the teacher's persistent classroom practice over each year, based on the variation in the 4-window average scores across teachers (between-teacher variance) and the variation in a teacher's scores across the windows (within-teacher variance). (See appendix C for details on the

⁵⁶ For each year, the distributions of CLASS and FFT overall scores in each window and the averaged scores across the four windows are presented in appendix exhibits D.7a and b, respectively.

⁵⁷ For mean overall scores by window, see Exhibit D.7c. The means are based on ratings from both principals and SHOs. There was no consistent difference between the scores of principals and study-hired observers. For CLASS, the average score from study-hired observers was higher than those from administrators in Year 1, but not in Year 2. For FFT, the average score from study-hired observers was higher than those from administrators in Year 2, but not in Year 1. For detailed results about differences in ratings between the two types of observers, see appendix exhibit D.8.

⁵⁸ As described in chapter 1, we also used the FFT and CLASS to code video-recordings of a sample of lessons for both treatment and control teachers in the spring of Year 2, to assess the impact of the intervention. The distribution of ratings based on the video-recordings for treatment teachers, shown in appendix exhibit D.7d, can be compared graphically with the distributions based on the intervention observers for Year 2, shown in exhibits D.7a and b. Appendix exhibits D.9a and b present statistical tests of the difference in means for the video and intervention score. The mean based on the video-recorded lessons is lower than the intervention ratings (4.63 for CLASS, compared to 5.54 for the Year 2 intervention ratings; 2.61 for FFT, compared to 3.08 for the Year 2 intervention ratings). The fact that the video scores were lower might be a result of the differing methods (live observation versus video-based coding), or the differing intended uses for the ratings (feedback versus analysis by the study team), or other factors such as the methods for supervising and supporting the raters.

⁵⁹ Classroom practice ratings from a single observation could also inform feedback about a teacher's instruction during a particular lesson, even if that performance were not indicative of a teacher's general instruction over the year. We do not have the necessary data to estimate the reliability of using single observations for feedback about instruction specific to a given lesson.

⁶⁰ The Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education 2014) do not suggest a minimum degree of reliability, but state that the reliability evidence for a measure should be appropriate for the measure's intended use, and that 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), although job performance ratings have often been found to have reliabilities below .70 (Viswesvaran, Ones, and Schmidt 1996).

estimation varied methods and results.) These reliability estimates tell us how consistent a teacher's overall scores were over the four observation windows.

The results of the reliability analyses for Years 1 and 2 are similar.⁶¹

- The 4-window average overall scores contained measurement error but provided some reliable information about a teacher's classroom practice over each year. In Year 2, for example, depending on assumptions about the sources of variation in the ratings each teacher received, reliability estimates for the 4-window average overall scores were between .53 and .61 for CLASS and between .70 and .77 for FFT. The reliability estimates for Year 1 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 each year because of variation in a teacher's overall scores across the four observation windows. In Year 1 and Year 2, the reliability of overall scores based on a single observation was .24 and .33 for CLASS, respectively, and .49 and .51 for FFT, respectively. In other words, 24 to 33 percent of the variation in CLASS overall scores and 49 to 51 percent of the variation in FFT overall scores represented between-teacher differences in classroom practice.

These reliability estimates are based on the assumption that all of the variation in a teacher's performance from window to window is due to idiosyncratic factors. To the extent that teachers improved over time in response to feedback, as implied by the theory of action, treating all variation across windows as measurement error would underestimate the true reliability of the observations. While we found statistically significant improvement over the four observation windows (see appendix exhibit D.7c), improvement can explain only a small portion of the variation in a teacher's performance across the windows. Thus, taking it into account would lead to only a small increase in reliability.⁶²

As another way of assessing reliability, we examined whether a teacher's 4-window average score in Year 1 was associated with the teacher's 4-window average score in Year 2. The correlation between these two scores was 0.51 for CLASS and 0.61 for FFT. These correlations are further evidence that the 4-window average scores provide some reliable information about a teacher's persistent classroom practice.⁶³

⁶¹ 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, for a 4-window average, Casabianca et al. (2013) reported reliabilities for CLASS that range from .32 to .72, and Ho and Kane (2013) reported reliability for FFT of .66.

⁶² For CLASS, average improvement in Year 2 ratings over the year potentially accounts for 10 percent of the measurement error (within-teacher variation). For FFT, average improvement in Year 2 ratings over the year potentially accounts for 4 percent of the measurement error. See appendix C, page C-5.

⁶³ In addition to examining the reliability of the classroom observations provided as part of the intervention, we also examined one aspect of their validity: their correlation with teachers' value-added. We estimated the correlation of scores from the Year 2 intervention observations with teachers' prior year VAM scores. (We used prior-year value-added 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.) The results appear

While most teachers received ratings that differed across dimensions, the scores did not provide a consistent message about which dimension the teacher most needed to improve. The dimension scores a teacher received in an observation report typically spanned two or more performance levels (43 to 74 percent for CLASS and 68 to 82 percent for FFT), indicating stronger performance on some dimensions and weaker performance on others.⁶⁴ However, the dimension scores did not convey a consistent message across observations about a teacher’s relative performance on the dimensions. In Year 2, for example, just 26 percent of teachers for CLASS and 34 percent for FFT had the same lowest-scored dimension of classroom practice for each of the four observations. Consistent with these results, the reliability of the difference between two different dimensions in a single observation was .19 for CLASS in both years and .09 and .12 for FFT in Years 1 and 2, respectively. Averaging across the four windows, the differences between dimension scores still had limited reliability to identify what a teacher most needed to improve: depending on assumptions about the sources of variance, .35 to .43 for CLASS and .18 to .30 for FFT. (See appendix C for details about the estimation methods and appendix exhibit C.1 for all reliability estimates.)

In addition to ratings, the reports prepared by observers were supposed to contain narrative text. 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 from Year 1.⁶⁵ Three-quarters of the sampled CLASS reports supported the dimension(s) of practice identified for improvement with at least one example from the observation. Less than one-quarter (23 percent) of the sampled FFT reports did so. These results might reflect the difference in reporting requirements between CLASS (which required observers to fill out all fields) and FFT (which did not require observers to fill out all dimension-specific fields).

Perceptions of the Performance Information on Classroom Practice

The intervention was intended to provide educators with performance information that was clearer, fairer, and more useful as a guide for professional growth than the information that they normally receive from the district. 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 hard to understand or not sufficiently specific, for example—they might ignore it and continue their normal classroom practices. For this reason, the survey administered in Year 2 asked treatment teachers to compare the study’s feedback on classroom practice to the feedback they received before the study. Seven items focused on whether the study’s feedback was better or worse, and one focused on how critical the feedback was of the teacher’s performance.

in appendix exhibit D.7e. The results indicate statistically significant positive correlations of the four-window average observation score and overall value-added (0.18 for CLASS and 0.31 for FFT). For comparison, we also computed the correlation of the scores based on video-recorded lessons and value-added; the correlation with overall value-added is statistically significant for CLASS (0.25) but not for FFT (0.10).

⁶⁴ For detailed results, see appendix exhibit D.10. See also appendix exhibits D.9a and b, which report the means by dimension on the intervention observations for treatment teachers in CLASS and FFT districts, as well as the means by dimension on the video-recorded lessons used in assessing the impact of the intervention.

⁶⁵ The analysis of the narrative content of the reports was performed using only Year 1 reports.

Three-quarters of treatment teachers said that the study's feedback on classroom practice was better than previous feedback from the district, averaging over seven characteristics of the feedback. For example, 65 percent of teachers reported that the study's feedback was more useful than the district's feedback, and 79 percent reported that it was more specific on what constitutes high-quality teaching. (See exhibit 2.4; for separate results for CLASS and FFT districts, see appendix exhibit D.11.)

The study's feedback gave almost all teachers the highest or second-highest possible performance level, so one might expect teachers to view the feedback as less critical than the district's feedback. However, about two-thirds (68 percent) of the treatment teachers in Year 2 reported that the study's feedback was "more critical of my performance" than the district's feedback.

Principals served as observers and had access to classroom observation reports created by the study-hired observers. Although they were not asked to compare study and district feedback, they were asked for their views on the study feedback. Nearly all of the treatment principals held positive views. In Year 2, for example, almost all (≥ 95 percent) reported that the CLASS/FFT system did a good job of distinguishing effective from ineffective teaching. (See exhibit 2.4; for separate results for CLASS and FFT districts, see appendix exhibit D.12. For results from Year 1, see chapter 2 of Wayne et al. 2016.)

Exhibit 2.4. Percentage of treatment teachers and principals who agreed somewhat or strongly with each statement about the feedback they received from the study's CLASS/FFT observations, Year 2

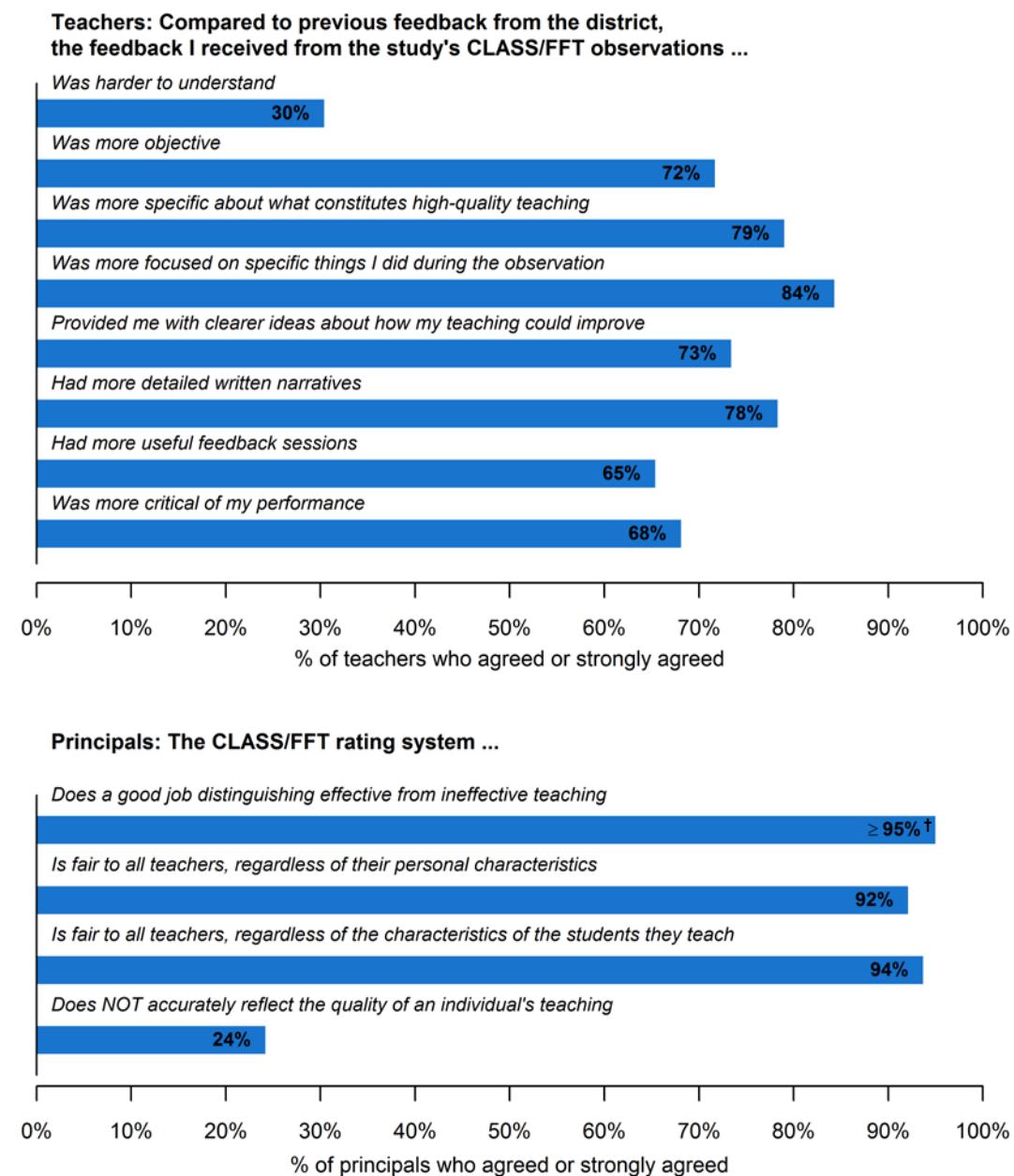


EXHIBIT READS: Of treatment teachers in Year 2, 30 percent agreed somewhat or strongly with the statement "The feedback I received from the study's CLASS/FFT observations was harder to understand" compared with the feedback received prior to the intervention as part of their district's approach to formal evaluation.

NOTES: Teacher sample size = 320–430 teachers; principal sample size = 60 principals.

† Reporting standards not met, too few cases to report the exact percentage.

SOURCES: Spring 2014 Teacher Survey and Spring 2014 Principal Survey.

The Intervention's Measure of Student Growth

The intervention's measure of student growth was intended to differentiate teacher performance to identify lower- and higher-performing teachers. In addition, the measure is intended to provide information about a teacher's relative performance in reading/ELA and mathematics, for those who taught both subjects. This section begins by describing the design of the student growth measure and discussing findings about how fully it was implemented. The section then examines how well the measure differentiated teacher performance. It concludes with findings on teachers' and principals' perceptions of the study's feedback on student growth.

Overview of the Measure

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 using a statistical method for analyzing multiple years of students' test score data. (See appendix E 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).⁶⁶

AIR prepared three waves of student growth reports. (See exhibit 2.5.) The first wave of reports was released between February and April of Year 1, prior to the spring surveys. The second and third waves were released in the fall of Year 2 and the fall of the year after the study.

Value-added scores can fluctuate significantly from year to year (Goldhaber and Hansen 2013; McCaffrey et al. 2009). For this reason, the reports showed a two-year average score, using value-added scores for the two previous years. Single-year value-added scores were reported for teachers who had scores for only one of the two previous years.

Exhibit 2.5. 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 (i.e., reading/ELA and mathematics). Each report presented a teacher's overall value-added score (i.e., average value-added score

⁶⁶ 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 the value-added score as a measure of student growth.

across grades and subjects), the score for each subject the teacher taught, and the score for each subject-grade combination. These scores were reported in student-level standard deviation units. To help readers compare a teacher's scores with other teachers' scores, the report presented the percentile rank for the teacher's 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 for his or her percentile rank.⁶⁷ (See appendix K for a sample student growth report for teachers.)

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

Implementation of the Intervention's Measure of Student Growth

To analyze the extent to which the student growth measure was implemented as intended, we examined how many teachers could be linked with a sufficient number of students to produce a value-added score, whether teachers and principals participated in training related to the student growth reports, and whether teachers and principals accessed or received 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. For each wave, there were sufficient data to estimate value-added scores for about 80 percent of teachers in total. For about two-thirds of teachers, those scores were based on two years of data. (See appendix exhibit F.1.)

In each wave, the implementation team made student growth reports available through a secure, web-based portal. Additional dissemination efforts for Wave 1 and Wave 2 differed.

Less than half of the teachers and principals accessed the Wave 1 student growth reports in Year 1, with access rates varying substantially across schools. Just prior to the release of the Wave 1 reports, the 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. These sessions were also intended to encourage teachers and principals to access their reports. Overall, 85 percent of teachers and 81 percent of principals participated in the webinars. After making the reports available through the portal, the team monitored access rates using the online portal and sent reminder notices to teachers who had not yet accessed their reports. Despite good attendance at the webinars, access rates were low:

⁶⁷ 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.

39 percent of the teachers with value-added scores and 40 percent of the principals accessed the reports online.^{68,69}

To ensure that all teachers saw the Wave 1 reports eventually, the team handed out hard copies at the classroom practice refresher training at the end of summer 2013. Each packet contained the teacher's Wave 1 report and most recent classroom practice report.

In fall of Year 2, hard copies of the Wave 2 student growth reports were viewed by all principals and received by 98 percent of teachers. Instead of encouraging teachers and principals to view their Wave 2 reports online, the implementation team convened an in-person workshop in each district at which principals were handed hard copies. Each principal was given a school-level report and a packet for each teacher that contained the teacher's most recent student growth and classroom practice reports. At the workshop, the team trained principals in how to interpret the reports. Principals were instructed to distribute the Wave 2 student growth reports directly to teachers in the weeks following the workshop. Each report packet contained a card for the teacher to send back to AIR to confirm receipt. AIR received cards from 98 percent of the teachers for whom the team was able to provide a report.

Principals were not required to have a feedback session with each teacher to discuss his or her Wave 2 report. However, in spring of Year 2, a little under half (47 percent) of the teachers with student growth reports reported that they had discussed their Wave 2 report with their principal.^{70,71}

Performance Information on Student Growth

In this subsection, we report on the potential utility of the student growth reports by first describing how many teachers' reports signaled that teachers needed to improve (or had excelled) and then assessing whether the information was reliable. In addition, for teachers who received value-added scores in both reading/ELA and mathematics, we discuss whether their reports suggested that their value-added was lower in one subject than the other, which could help teachers focus their improvement efforts.

In Wave 1 and Wave 2, just under a quarter of teachers with value-added scores in reading/ELA—and about half with value-added scores in mathematics—had a value-added score that measurably differed from the district average, thus

⁶⁸ 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.

⁶⁹ The 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 (with district averages ranging from 17 to 78 percent) and principals (with district averages ranging from 11 to 100 percent). For more information, see the study's first report (i.e., Wayne et al. 2016).

⁷⁰ This figure is based on a survey item that was administered only in Year 2.

⁷¹ The denominator for the percentage that appears in the text is all teachers with a student growth report. If we instead use all treatment teachers as the denominator, 37 percent reported that they had discussed their Wave 2 student growth report with their principal.

signaling that the teacher needed to improve or had excelled. As with all value-added measures, uncertainty in a teacher's value-added score means that 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 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 associated with 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 (type I error) and (2) the risk of misidentifying teachers who were truly below or above average as average teachers (type II error).⁷² Taking 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 2.6.)⁷³ For example, 12 percent of teachers with a reading/ELA value-added score had value-added scores that were measurably higher than the district average. In total, 23 percent of teachers in Wave 1 and 21 percent in Wave 2 had reading/ELA value-added scores that were measurably different from the average. Based on the mathematics value-added scores, 52 percent of teachers in Wave 1 and 47 percent in Wave 2 had a value-added score that was considered measurably different from the district average.⁷⁴

The value-added scores provided some reliable information for distinguishing between lower- and higher-performing teachers. In order to 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 that a teacher's value-added score should reflect persistent performance rather than idiosyncratic factors introduced by classroom composition or abnormal events. We estimated the degree to which the value-added scores were a reliable measure of persistent performance based on how much a teacher's score varied across the two years of student growth data, identified in exhibit 2.5. (See appendix C for details on the methods and results, and appendix exhibit C.1 for all reliability estimates.) 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 reliability for Wave 1 value-added scores was estimated to be .44 for reading/ELA and .68 for mathematics. For Wave 2, the reliabilities were .46 and .67, respectively. These reliability estimates are consistent with

⁷² 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, Raudenbush and Jean (2012) discussed the tradeoff between a 95 and 75 percent confidence interval, noting that teachers might wish to use the latter for self-evaluation.

⁷³ 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.

⁷⁴ 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, then fewer treatment teachers would be considered measurably above/below average. For overall value-added scores in Year 2, 80 percent would not be measurably different from average, 13 percent would be measurably above average, and 7 percent would be measurably below average.

estimates found in research on other value-added measures (Goldhaber and Hansen 2013; McCaffrey et al. 2009; Mihaly et al. 2013; Whitehurst, Chingos, and Lindquist 2014).^{75,76}

As discussed earlier, the student growth reports show teachers whether they improved student performance “measurably” more than or “measurably” less than a teacher with their district average score, or whether they were indistinguishable from a teacher with their district average score. Among teachers with both Wave 1 and Wave 2 value-added scores for reading/ELA, 74 percent remained in the same classification across waves, demonstrating a consistent message about performance from one year to the next. For mathematics, 69 percent of teachers remained in the same classification.⁷⁷ A high degree of consistency from one wave to the next was expected because a teacher’s value-added score for each wave was a two-year average, when sufficient data were available.

⁷⁵ Other studies tend to report year-to-year correlations in value-added scores between .30 and .67, which implies the reliability of value-added scores based on two years of data are typically between .46 and .80. Value-added scores for mathematics and middle school teachers are generally more reliable than value-added scores for reading/ELA and elementary school teachers.

⁷⁶ For a discussion about the year-to-year variability in value-added scores, see Raudenbush and Jean (2012).

⁷⁷ Of teachers with a value-added score for reading/ELA for Wave 2, 78 percent had a score for Wave 1 as well. For mathematics, 81 percent of those with a mathematics value-added score in Wave 2 also had one for Wave 1.

Exhibit 2.6. Distribution of treatment teachers based on whether their value-added score in each wave was considered measurably above or below the district average, overall and by subject

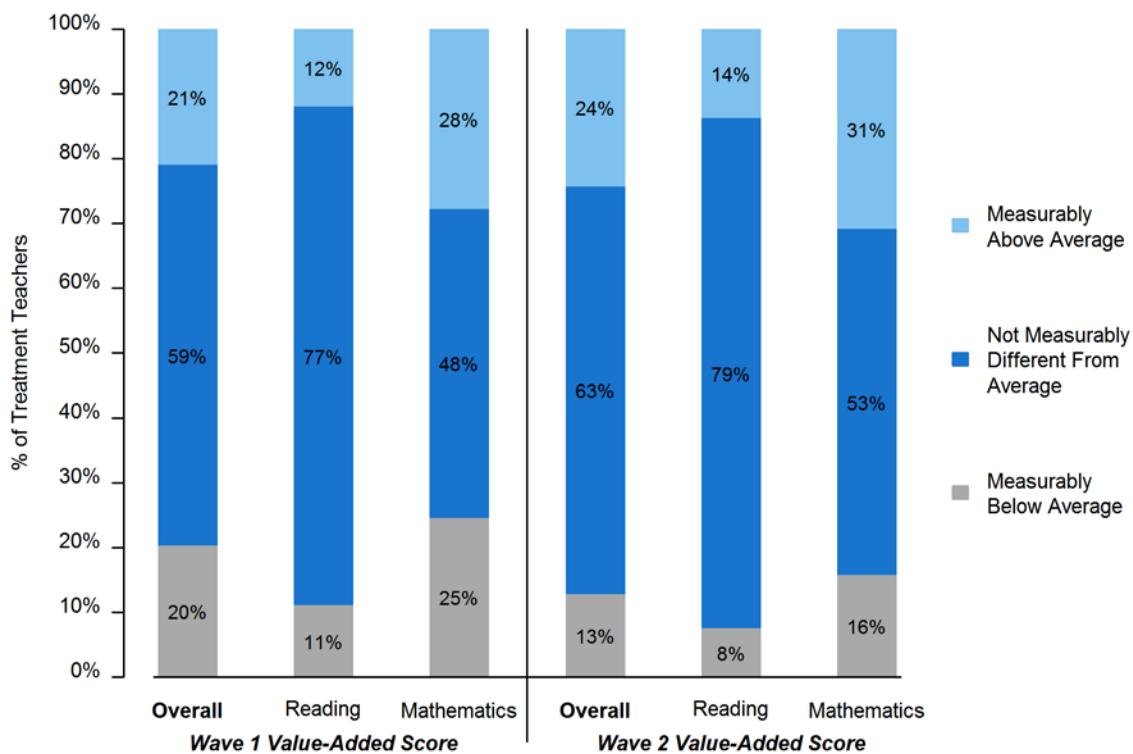


EXHIBIT READS: For treatment teachers with Wave 1 overall value-added scores, 21 percent had scores considered measurably above the district average.

NOTES: 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 for Year 1 = 433 teachers with overall value-added scores; 326 teachers with reading/ELA value-added scores; and 342 teachers with mathematics value-added scores. Sample size for Year 2 = 415 teachers with overall value-added scores; 320 teachers with reading/ELA value-added scores; and 330 teachers with mathematics value-added scores.

Reported percentages may not sum to 100 percent because of rounding.

SOURCE: AIR Value-Added system.

Among teachers with value-added scores in both reading/ELA and mathematics, about half had student growth reports that suggested the teacher performed better in one subject area than the other. Of the teachers with value-added scores, a little over half (55 percent in Year 1 and 57 percent in Year 2) 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, the teachers could infer whether their performance in each subject was measurably below average, not measurably different from average, or measurably above average. In total, a little under half of teachers with scores in both subjects had student growth reports that suggested different

performance in reading/ELA than mathematics (48 percent in Year 1 and 41 percent in Year 2).^{78,79} (See appendix exhibit F.2.)

Perceptions of the Performance Information on Student Growth

In Year 2, we asked treatment teachers and principals for their views about the feedback on student growth because those views might affect how they reacted to the feedback.⁸⁰ If teachers or principals had negative views of the feedback, seeing it as unfair, for example, they might ignore it.

Only about half of the teachers and three-quarters of the principals were positive about the student growth reports they received. About half of treatment teachers (41 to 55 percent) reported positive views about the study's feedback on student growth. For example, 48 percent of teachers reported that "the value-added score is a good measure of how well students learned what I taught last year." Likewise, about 42 percent of teachers reported that the value-added scores are fair to all teachers, regardless of the personal characteristics of the students they taught. For principals, the percentages responding that value-added scores were a good measure of how well students learned and fair to all teachers regardless of students' personal characteristics were 74 and 75 percent, respectively. (See exhibit 2.7; for details, see appendix exhibits F.2 and F.3.)

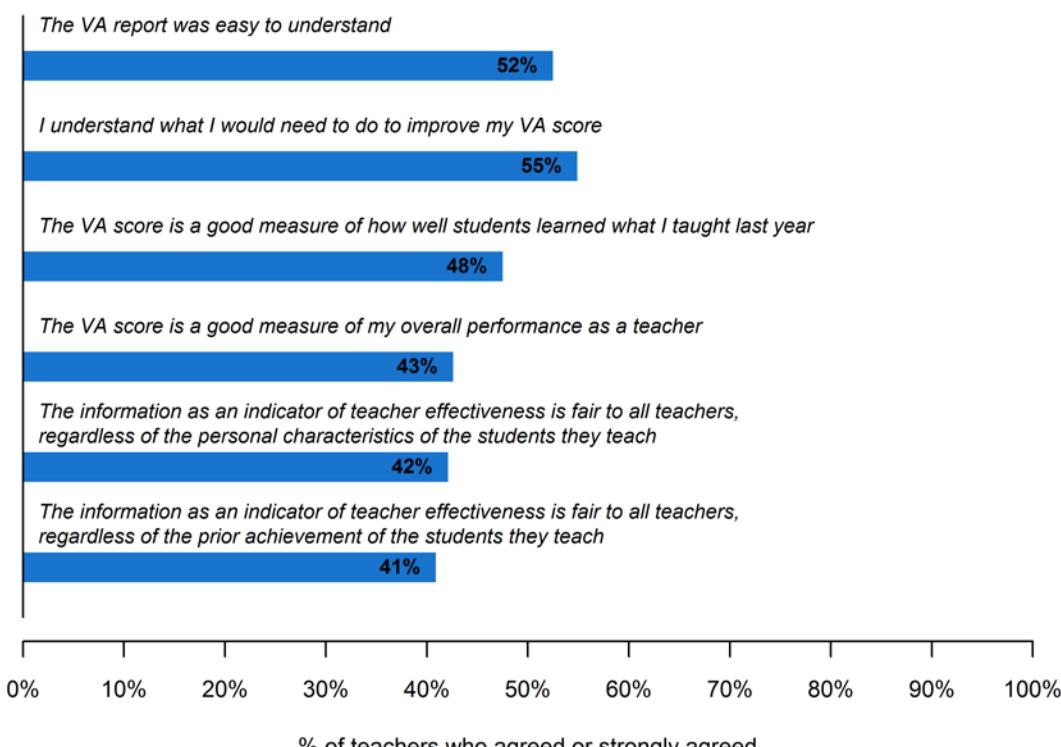
⁷⁸ We examined differences in a teacher's subject-specific value-added scores, which are based on student growth in test score standard deviation units in each subject. The student growth reports also included the teacher's value-added percentile ranking in each subject. We based the analysis on the 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.

⁷⁹ We also estimated the degree to which the difference between a teacher's value-added scores in reading/ELA and mathematics was 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 for Wave 1 and .50 for Wave 2. (See appendix C for details about the estimation method and results.)

⁸⁰ The findings discussed in the next paragraph are based on survey items that were included only in the Year 2 surveys and were administered to respondents only in the treatment group. The Year 1 surveys asked about educators' perceptions of the feedback they had received but did not ask for perceptions of the intervention's feedback specifically. See chapter 5 of Wayne et al. (2016).

Exhibit 2.7. Percentage of treatment teachers and principals who agreed somewhat or strongly with statements about the student growth reports they viewed, Year 2

Teachers



Principals

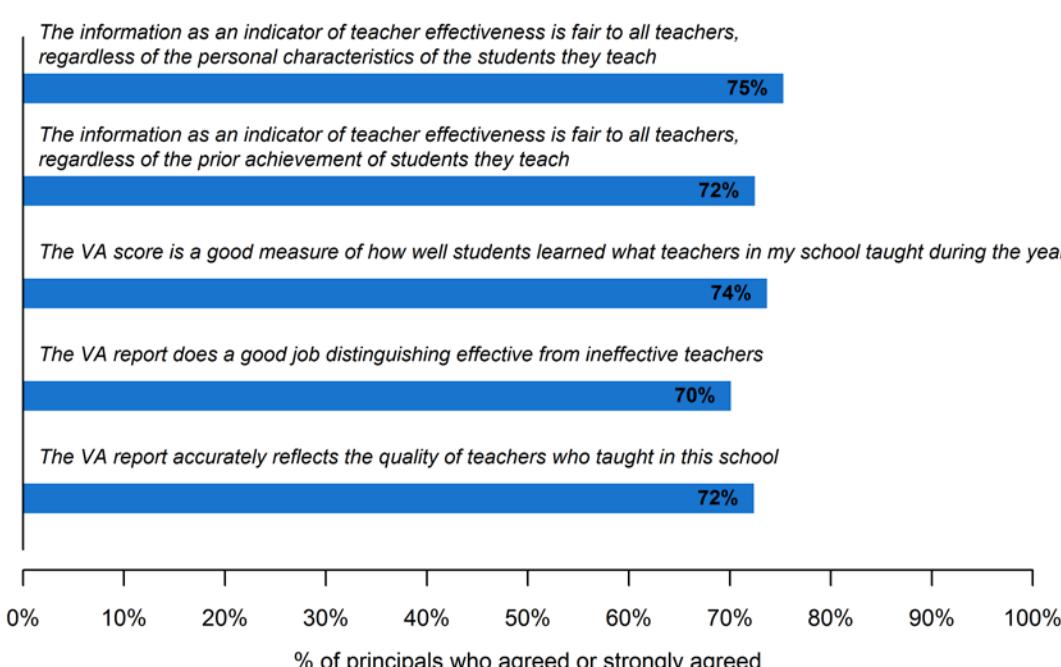


EXHIBIT READS: Of treatment teachers in Year 2 who reviewed their student growth report, 52 percent agreed somewhat or strongly with the statement "The VA report was easy to understand."

NOTES: Teacher sample size = 311–315 teachers; principal sample size = 51 or 52 principals.

SOURCES: Spring 2014 Teacher Survey and Spring 2014 Principal Survey.

The Intervention's Measure of Principal Leadership

This section presents findings about the intervention's measure of principal leadership as implemented. The information provided as feedback by the measure, the Vanderbilt Assessment of Leadership in Education (VAL-ED), was intended to identify lower- and higher-performing principals and potentially identify principals who need additional support. In addition, the measure was intended to identify practices that, if improved, would lead to more effective leadership and higher student achievement.

Overview of the Measure

The VAL-ED is a 360-degree survey of principals, their supervisors, and teachers. The VAL-ED was used to provide principals and their supervisors with information about principal leadership behaviors associated with student learning, and it was administered in fall and spring of both study years. The dimensions measured include six “core components” and six “key processes.” (See exhibit 2.8; see appendix exhibit G.1 for definitions of the core components and key processes.) In addition, it measures 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 2.8. VAL-ED core components and key processes

Core components	Key processes
<ul style="list-style-type: none">• High standards for student learning• Rigorous curriculum• Quality instruction• Culture of learning and professional behavior• Connections to external communities• Systemic performance accountability	<ul style="list-style-type: none">• Planning• Implementing• Supporting• Advocating• Communicating• Monitoring

The VAL-ED survey asks each respondent to use a 5-point scale to rate a principal’s effectiveness in 72 leadership behaviors that represent the 36 component-by-process areas.⁸¹ (See appendix exhibit G.2 for a sample of survey items.) The online system collects the responses electronically and produces a report on the principal. 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. Scores are also reported separately by respondent group. (See appendix exhibit G.4.)

To aid principals and their supervisors in interpreting the ratings, the developer assigned each score a performance level (*below basic*, *basic*, *proficient*, or *distinguished*).⁸² (See appendix

⁸¹ In both fall and spring, each principal and 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.

⁸² The developer used a standard-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).

exhibit G.3 for a screen shot from a sample VAL-ED report that includes the performance-level descriptors.) The report also presents percentile ranks corresponding to each score based on how the principal performed relative to the principals in a national VAL-ED field test.

The VAL-ED report also presents the score for each component-by-process combination in a six-by-six matrix, with color-coded cells indicating performance level. (See appendix exhibit G.5.) 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.”

Implementation of the Intervention’s Measure of Principal Leadership

This subsection 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.

The VAL-ED training for principal supervisors was designed to provide principals with structured feedback, using the report as the focus of a feedback session. 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 should grow. All principals and their supervisors participated in a two-hour training, held initially in the summer prior to the first study year (i.e., summer 2012). In addition, all principal supervisors received training before each wave of feedback sessions. Training for the fall Year 1 session was designed to last up to one hour; for subsequent sessions, it could be shortened.

All VAL-ED reports in both study years incorporated input from the principal, the principal’s supervisor, and most teachers. The average teacher response rates at each school for the four VAL-ED administrations were 80 percent, 90 percent, 86 percent, and 88 percent, respectively.

Feedback sessions generally occurred as planned. After each VAL-ED administration, nearly all principals met with their supervisors to discuss their reports.⁸³ In Year 1, the supervisors reported that the feedback sessions lasted 52 minutes, on average, in the fall and 46 minutes in the spring. In Year 2, the sessions lasted 36 minutes in the fall and 34 minutes in the spring. We did not ask principals to report on the duration of feedback provided by the intervention.

Performance Information on Principal Leadership

As described earlier in this section, the principal leadership measure provided detailed information for principals on their leadership. In this subsection, we focus first on the overall performance levels, examining the distribution of principals across performance levels within each assessment window (fall and spring) each year. We then examine whether the measure was sufficiently reliable to identify principals in need of support. Finally, we discuss how well the principals’ reports identified the dimensions of leadership that principals needed to work on the most.

⁸³ In each of the two study years, each principal in a treatment school participated in at least one feedback session. In Year 2, a small number of principals did not participate in a second feedback session.

In all four administrations, principals were distributed across all four performance levels, with many principals receiving ratings that indicated room for improvement.

This contrasts with the classroom practice ratings, almost none of which were below the highest two performance levels. Across the four VAL-ED administrations, from 41 percent (spring Year 2) to 70 percent (fall Year 1) of principals had overall scores in the bottom two performance levels (*below basic* and *basic*). (See exhibit 2.9.)

Exhibit 2.9. Distribution of treatment principals across performance levels based on VAL-ED overall scores in fall and spring, by year

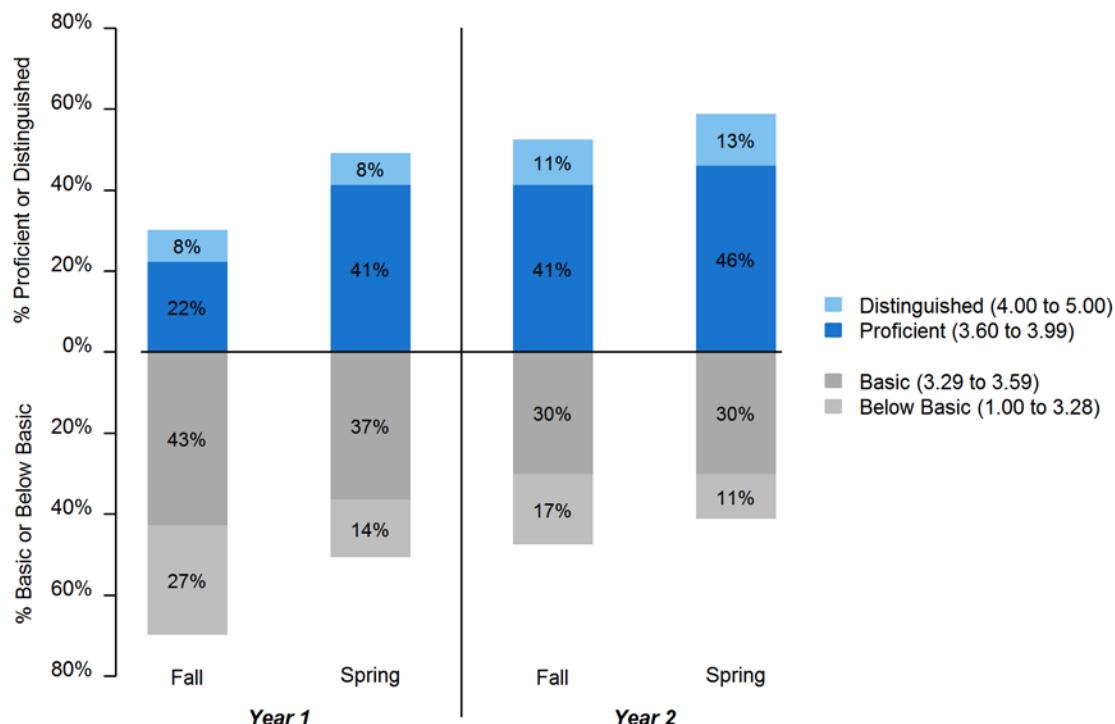


EXHIBIT READS: In fall of Year 1, 8 percent of treatment principals had a VAL-ED overall score at the *distinguished* level, 22 percent at the *proficient* level, 43 percent at the *basic* level, and 27 percent at the *below basic* level.

NOTES: 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. Reported percentages may not sum to 100 percent because of rounding. Sample size = 63 principals for Year 1 (fall 2012, spring 2013) and Year 2 (fall 2013 and spring 2014).

SOURCES: Fall 2012, Spring 2013, Fall 2013, and Spring 2014 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. Based on reports in the fall of Year 1, the mean score was at the 37th percentile. In the spring of Year 2, the mean score was at the 56th percentile.⁸⁴ The change in national percentile ranking over time does not necessarily mean

⁸⁴ The mean scores shown in the text are based on the principals of all 63 schools. The mean VAL-ED score for the 50 principals who were present in the fall of Year 1 and the spring of Year 2 increased from the 37th percentile in the fall to the 55th percentile in the spring, a statistically significant change. See appendix exhibit G.6 for mean overall VAL-ED scores for each of the four administrations, and see appendix exhibit G.7 for mean scores by rater type for the four administrations.

principal leadership improved. Respondents who are asked to complete the VAL-ED surveys multiple times for the same principal may be inclined to give higher ratings each time; the national norming sample upon which the percentile rankings are based does not account for principals receiving multiple ratings over time.

VAL-ED ratings provided by principals, supervisors, and teachers in the fall administrations were often too different to form a reliable measure, but the spring ratings were consistent enough to identify some principals as needing support.

To provide information on a principal's overall effectiveness, the VAL-ED scores from each of the three types of raters should communicate a consistent message about the principal's effectiveness. Based on the literature on 360-degree surveys, we would expect correlations of .25 to .35 between respondent group scores.⁸⁵ In each fall administration, however, agreement among the three respondent groups' overall scores was low, with two of the three correlations below .10. (See exhibit 2.10.) In the spring administrations, correlations were higher, and thus the reports provided a more consistent message about a principal's effectiveness. We do not have evidence to explain why the correlations generally were higher in the spring than the fall.⁸⁶ We estimated that the VAL-ED overall score reliability (i.e., inter-rater reliability) was .19 and .32 in the two fall administrations and .51 and .49 in the two spring administrations. (See appendix C for details on the estimation methods and results.)⁸⁷ The improved reliability in spring reflects greater average agreement among the respondent groups.⁸⁸

⁸⁵ For the VAL-ED correlations, see Porter et al. (2010). For the literature on 360-degree surveys, see Conway and Huffcutt (1997).

⁸⁶ The correlations between scores given by the three different types of respondents, discussed above, indicate the extent to which principals who received relatively high ratings from one type of rater (e.g., the supervisor) also received relatively high ratings from other types of raters (e.g., teachers). We also examined whether the respondent groups differed in the average ratings they provided. The patterns we found were not consistent from fall to spring. In the fall of both years, average overall scores were similar across the three respondent groups (teachers, supervisors, and principals). But in the spring there were some statistically significant differences. Principal self-ratings were higher than teacher ratings in spring for both years. Also, in spring Year 2, supervisor ratings were higher than teacher ratings. (For detailed comparisons, see appendix exhibit G.7.)

⁸⁷ As a point of reference, reliability for the classroom observation 4-window average scores in Year 1 was estimated to be between .42 and .75.

⁸⁸ In addition to examining the reliability of the overall scores based on three respondent groups, we also examined the reliability of the ratings given by teachers. 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. Because 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.

Exhibit 2.10. Correlations between VAL-ED respondent group overall scores from different respondent groups in fall and spring, by year

Correlation	Fall Year 1	Spring Year 1
Principal and supervisor	.08	.27*
Principal and teachers	.06	.26*
Supervisor and teachers	.27*	.38*
Correlation	Fall Year 2	Spring Year 2
Principal and supervisor	.01	.23
Principal and teachers	.34*	.28*
Supervisor and teachers	.08	.38*

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

NOTES: Sample size = 63 principals for fall 2012, spring 2013, fall 2013, and spring 2014.

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

SOURCES: Fall 2012, Spring 2013, Fall 2013, and Spring 2014 VAL-ED Surveys.

We examined whether the principals who stayed in their schools during the two-year study received a consistent message about their performance across the two years.⁸⁹ (A principal who receives inconsistent messages across years might not accept the feedback as valid, limiting its effect.) We found that the messages principals received were reasonably consistent. Based on the overall scores computed for each year, 55 percent of principals present in both years had the same performance level in both years. The remaining principals had performance levels that differed across years by one level: 10 percent had a lower performance level in Year 2 than Year 1, and 35 percent had a higher level.

Almost all reports showed dimension scores that spanned multiple performance levels, but these scores did not reliably indicate which dimension a principal most needed to work on. To inform decisions about improving practice and to identify professional development needs, each VAL-ED report provided 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). If a principal received different ratings on different aspects of his or her 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. Nearly all principals received scores that differed across different dimensions of principal leadership.⁹⁰ (For detailed results, see appendix exhibit G.8.) However, 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.

To analyze the consistency in a principal's dimension scores across respondent groups, we examined the degree to which a principal's dimension scores from the three respondent groups formed a reliable measure of whether a principal's performance was better in some dimensions

⁸⁹ Fifty-one principals in the treatment schools received VAL-ED reports in both years of the study.

⁹⁰ The percentage of principals who received dimension scores that did not differ is not provided to protect data confidentiality.

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.) Across the four administrations, we estimated that the reliability of the difference between two of a principal's dimension scores ranged from .36 to .50 for the core components and .07 to .31 for the key processes. Thus, the dimension scores did not reliably indicate which dimension a principal needed to work on the most. (See appendix C for details about the estimation methods and appendix exhibit C.1 for all reliability estimates.)

Perceptions of the Performance Information on Principal Leadership

Because a principal's views of the feedback received might affect how he or she responded to it, we asked principals a set of questions about the study's feedback on their leadership, similar to the questions we asked teachers.⁹¹ Five items focused on whether the study's feedback was better or worse than previous feedback from their district. In addition, one item focused on how critical the feedback was of the principal's performance.

On four of five items, nearly three-quarters of principals said the study's feedback on leadership was better than previous feedback from the district. For example, 73 percent reported that the feedback they received from the VAL-ED was more objective than the feedback they received from the district system, and 75 percent reported that it provided "clearer ideas about how to improve my leadership." (See exhibit 2.11, and for details see appendix exhibit G.9.) These results are similar to those for teachers' perceptions of the feedback on classroom practice. On one of the five items, however, over half (55 percent) of principals reported that the study's feedback was less "comprehensive" than previous feedback from the district, even though the study's feedback on leadership spanned many dimensions.

Because many principals received overall ratings in the bottom two of the four VAL-ED performance levels, principals might have seen the study's feedback on their leadership as more critical than their previous feedback from the district. Just over half (58 percent) of the principals said the study's feedback was "more critical of [their] performance" than the district's feedback prior to the study.

⁹¹ The findings in this subsection are based on survey items that appeared only in the Year 2 surveys and were administered to principals only in the treatment group. The Year 1 surveys asked about principals' perceptions of the feedback they had received on their leadership but did not ask for specific perceptions of the intervention's feedback. See chapter 5 of Wayne et al. (2016).

Exhibit 2.11. Percentage of treatment principals who agreed somewhat or strongly with statements about the feedback they received from the VAL-ED, Year 2

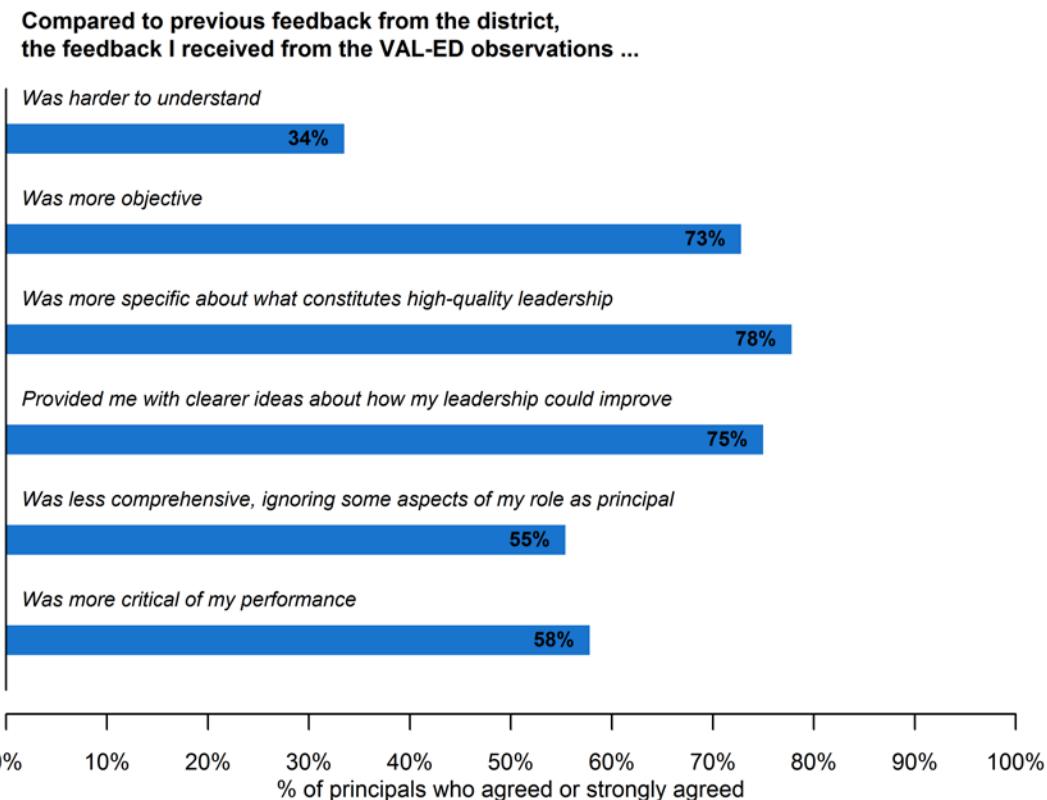


EXHIBIT READS: Of treatment principals in Year 2 who received feedback from the VAL-ED, 34 percent agreed somewhat or strongly with the statement "Relative to the district's approach to evaluation, the feedback I received from VAL-ED observations harder to understand."

NOTES: Sample size = 45 principals.

SOURCE: Spring 2014 Principal Survey.

Summary

The intervention provided feedback on classroom practice, student growth, and principal leadership. The feedback was intended to identify the educators who needed support and indicate the area of practice an educator most needed to improve.

Most teachers received the intended number of classroom observations in both years. Although very few received classroom practice ratings that signaled a need to improve, and scores for single observations were not very reliable, overall scores averaged across four observations provided some reliable information to identify teachers who most needed support. The scores for individual dimensions of classroom practice were not reliable enough to identify the area of practice that a teacher most needed to improve.

During Year 1, most teachers did not access their student growth reports, which were available online. However, printed copies of almost all reports were successfully delivered in Year 2.

Many teachers received reports indicating that they performed measurably above average, indicating they had excelled, or below average, signaling a need for improvement. In addition, the reports had the potential to signal which subject to focus on for improvement; about half of the teachers with scores in both ELA and mathematics received student growth reports that suggested they needed to improve more in one subject area than the other.

Finally, the feedback on principal leadership was gathered and delivered as planned and often indicated a need to improve. The ratings from the two spring administrations were reliable enough to identify the principals who needed support, although the ratings from the two fall administrations were not. Almost all reports showed dimension scores that spanned multiple performance levels, but these scores did not reliably indicate which area a principal most needed to work on.

Chapter 3. Impact of Performance Feedback on Teacher, Principal, and Student Outcomes

According to the theory of action in chapter 1, the performance feedback was expected to identify educators for support and signal specific areas of practice for improvement. Providing performance feedback to teachers and principals was expected to affect their interest in improving in areas included in the feedback measures, their participation in professional development in these areas, and their perceptions of their performance. By affecting these “initial outcomes,” the intervention would motivate improved performance, focus teachers’ and principals’ attention on practices that might be effective, and strengthen their knowledge and skills. This, in turn, would lead to improved teacher classroom practice and principal leadership, ultimately boosting student achievement.

This chapter examines whether the intervention had these anticipated effects. The chapter begins by examining whether the intervention generated the intended contrast between treatment and control schools in educators’ experience with feedback. The chapter then examines the impact of the intervention on educators’ initial outcomes. It concludes by discussing the impact on classroom practice, principal leadership, and achievement.

The intervention was implemented over two years, under the hypothesis that teachers and principals might continue to benefit from feedback even after several rounds. To assess this hypothesis, results are reported separately for Year 1 and 2. Each year, the analyses were based on all teachers and principals present in the spring, along with the students in the teachers’ classes.^{92,93} Thus, due to staff mobility, the Year 2 sample includes Year 1 teachers and principals who remained in their schools as well as those new to their schools. Because some staff and students were new in the second year, not all of those in the Year 2 impact sample in the treatment schools had the opportunity to experience two years of the intervention.⁹⁴ Although one might anticipate larger impacts in the second year, due to added opportunity for feedback, it is possible that turnover or other factors might limit the cumulative impact.

The main focus is on the average impact of providing performance feedback across the eight study districts. As described in chapter 1, the intervention used the CLASS observation measure to provide feedback on classroom practice in four of the eight districts, and the FFT in the other four. To examine the robustness of the main results, analyses were also conducted for these two sets of districts separately. The results on educators’ experiences and on initial outcomes in CLASS and FFT districts are referenced in footnotes; the results on the intervention’s impact on the study’s primary outcomes (classroom practice, principal leadership, and student

⁹² See appendix H for a description of the analysis methods and samples for the impact analyses.

⁹³ Analyses of teacher and student outcomes were based on grades 4–8, which were the main focus of the intervention. Teachers of kindergarten through grade 3 in treatment schools participated in some aspects of the intervention to promote schoolwide engagement (see chapter 1). 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. Results for teachers in grades K–3 corresponding to exhibits 3.2, 3.3, and 3.5 appear in appendix I.

⁹⁴ See appendix exhibits A.5 and 6 for an analysis of principal and teacher entries and exits over the period of the study, and exhibits A.7 and 8 for an analysis of student entries and exits.

achievement) in CLASS and FFT districts are presented within the main text. The separate results for the two sets of districts should be interpreted with caution because 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 impacts discussed in this chapter are statistically significant at the .05 level based on two-tailed tests.

Key Findings

Contrast in Educators' Experience With Feedback

- Each year, treatment teachers reported receiving substantially more feedback sessions with ratings and a written narrative on their classroom practice than control teachers (for example, 3.0 versus 0.2 sessions in Year 2), and they were more likely to report receiving value-added scores than were control teachers (81 versus 34 percent in Year 2).
- Treatment principals reported receiving twice as many feedback sessions with ratings each year as control principals (for example, 2.0 versus 1.0 sessions in Year 2).

Impact on Initial Outcomes

- The intervention had little impact on teachers' initial outcomes. A higher percentage of treatment than control teachers reported discussing at least one CLASS/FFT topic with someone giving them feedback (for example, 89 versus 78 percent in Year 2), but it had no impact on the percentage indicating that they would like to improve on CLASS/FFT topics or that their professional development covered these topics.
- The intervention had no impact on principals' initial outcomes. It did not affect the percentage of principals discussing at least one VAL-ED topic with someone giving them feedback, the percentage indicating that they wanted to improve on these topics, or the percentage reporting that their professional development covered the topics.

Impact on Classroom Practice, Leadership, and Achievement

- The intervention had a positive impact on classroom practice based on video-recorded lessons coded using the CLASS, moving teachers from the 50th to the 57th percentile, but it did not have an impact on classroom practice coded using the FFT. The impact occurred only in the districts that used the CLASS for feedback.
- In Year 1, the intervention had a positive impact on teacher-principal trust, one of the two measures of leadership examined, moving principals from the 50th to the 60th percentile; in Year 2, it had an impact on both instructional leadership and teacher-principal trust.
- At the end of the first year, the feedback had a small positive impact on student achievement in mathematics, equivalent to about four weeks of learning. At the end of the second year, the impact was similar in magnitude but not statistically significant. There was no impact in either year on students' reading/ELA achievement.

Contrast in Educators' Experience of Feedback

The intervention was designed to substantially increase the amount and quality of feedback received by teachers and principals in treatment schools. As described in chapter 1, the study's performance feedback was provided in addition to the performance feedback the districts provided through their established teacher and principal evaluation systems. In this section, we assess whether the frequency of the feedback received by teachers in the treatment schools (combining feedback from the study and their district's standard process) differed from the feedback received by teachers in the control schools who received feedback through the standard process only.

Feedback for Teachers

The intervention was designed to provide teachers with feedback that incorporated numerical ratings of their classroom practice and incorporated a narrative discussion of their teaching and areas for improvement. In addition, the intervention was expected to provide teachers with information on their contributions to growth in student achievement.

As expected, treatment teachers reported receiving more feedback with ratings than control teachers. Each year, more than 80 percent of treatment teachers reported receiving feedback with ratings, compared with less than half of the control teachers.⁹⁵ (See appendix exhibits I.1a and 1b.) The results were particularly pronounced for nonprobationary teachers.⁹⁶ Nonprobationary teachers in treatment schools were much more likely than nonprobationary teachers in control schools to receive feedback with ratings (87 versus 35 percent in Year 2).

In both years, treatment teachers also reported more than four times as many feedback sessions with ratings and a written narrative as control teachers. Some researchers have argued that feedback accompanied by narratives may be especially helpful in supporting teachers in improving their instruction. (See, for example, Rowan and Raudenbush 2016.) In both Year 1 and Year 2, the average treatment teacher reported receiving 3.0 feedback sessions that included ratings and a written narrative, compared with 0.7 for the average control teacher in Year 1 and 0.2 in Year 2. (See exhibit 3.1.) The total duration of all oral feedback sessions received was also substantially higher for treatment than control teachers—for example, 100 minutes in Year 2 for the average treatment teacher, compared with 25 minutes for the average control teacher.

⁹⁵ The data on feedback were based on a survey administered in the spring of each year, which asked teachers to report on every instance in which they were observed and received feedback that year, including evaluation-related observations as well as walkthroughs and informal observations (e.g., peer-to-peer observations).

⁹⁶ 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.

Exhibit 3.1. Number of feedback sessions with ratings and written narrative and duration of oral feedback that an average teacher reported receiving, by treatment status and year

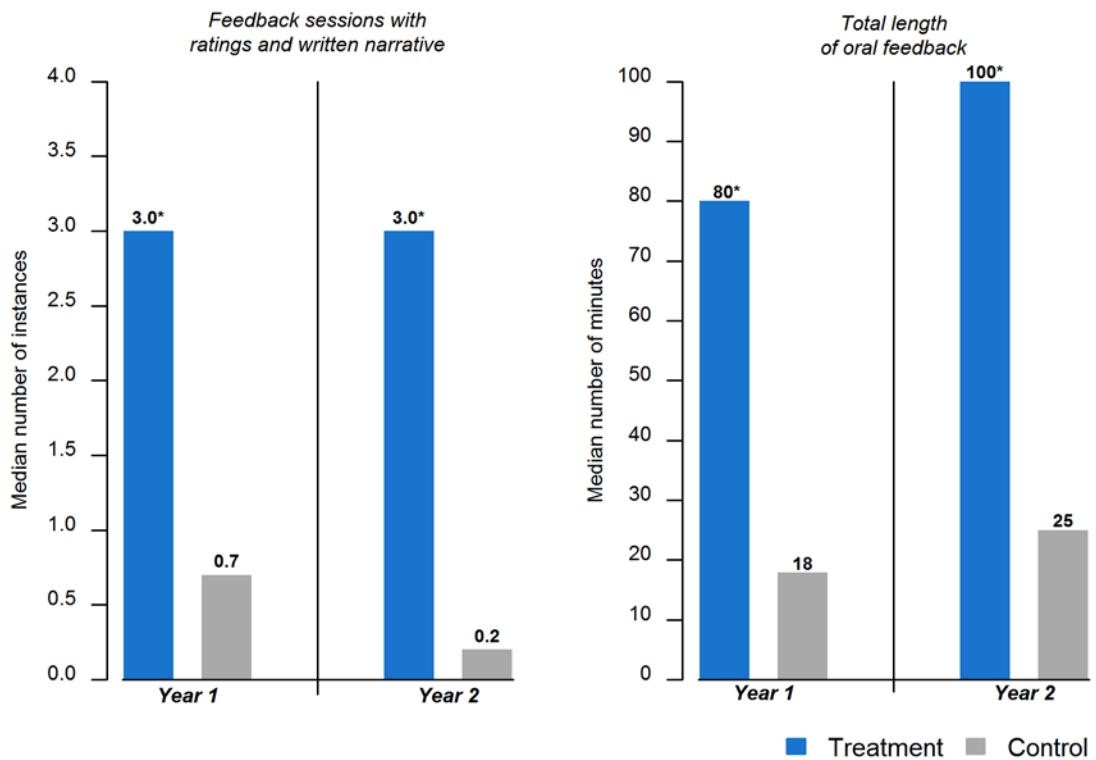


EXHIBIT READS: The average treatment teacher in Year 1 reported 3.0 feedback sessions with ratings and written narrative, compared with 0.7 for control teachers.

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

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

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

See appendix exhibits I.2a and 2b for separate results for CLASS and FFT districts and for grade K–3 teachers.

SOURCES: Spring 2013 and Spring 2014 Teacher Surveys.

In both years, treatment teachers were more likely than control teachers to report receiving feedback based on observations from observers not based at the teachers' schools. Drawing on evidence that ratings of classroom practice are more reliable if they are based on observations conducted by multiple observers, the intervention's measure of classroom practice was designed to provide teachers with observations by observers from outside their schools, as well as by their school administrator. As expected, the majority of teachers in both treatment and control schools reported being observed by a school administrator in both years. For example, in Year 2, three-quarters of the treatment and control teachers reported at least one observation conducted by their principal. However, there was a substantial treatment-control difference in the proportion of teachers reporting that they were observed by someone from outside their school (87 of treatment teachers in Year 2, compared with 15 percent for control teachers).⁹⁷ (See appendix exhibits I.2a and 2b for detailed results by year.)

⁹⁷ In the relevant item on the teacher survey, non-school-based observers excluded coaches or mentors.

In addition to feedback on their classroom practice, the intervention also provided teachers “student growth reports” that included a value-added score as a measure of each teacher’s contribution to student growth. As anticipated, a higher percentage of teachers in treatment than in control schools reported receiving value-added scores, especially in Year 2 (45 versus 24 percent in Year 1 and 81 versus 34 percent in Year 2). (See exhibit 3.2.)^{98,99} However, fewer treatment than control teachers reported receiving information on the *achievement of individual students* they taught (for example, 73 versus 88 percent in Year 2). It is not clear why fewer treatment teachers reported receiving information on individual students. Perhaps principals in treatment schools were less likely to distribute individual achievement results to teachers, knowing the teachers had access to growth reports; perhaps treatment teachers considered the value-added scores a substitute for data on individual students and thus were less likely to seek out such data.¹⁰⁰

⁹⁸ Although we lacked data to assess the validity of all teachers’ responses to the items about receiving student achievement information, we were able to examine the validity of treatment teachers’ responses in Year 1. Specifically, we compared treatment teachers’ responses on the spring Year 1 surveys with log-in records from the online system used to disseminate Wave 1 reports. We found some evidence that teachers may not have understood the distinction between value-added scores and other information about students’ achievement. About one-third (34 percent) of the treatment teachers who reported receiving value-added scores did not log-in, and 17 percent of those who reported not receiving value-added scores actually logged-in. (The Wave 2 reports were distributed in hardcopy, so we cannot examine validity in the same way using the Year 2 surveys.)

⁹⁹ The survey items asking teachers whether they received value-added information differed in Year 1 and 2. In Year 1, the item was included in a broader question asking about different types of achievement information. In Year 2, the survey included a separate question asking whether teachers received a value-added score representing the classes they taught. We made this change because of evidence that some teachers may not have understood the Year 1 item, as discussed in the previous footnote.

¹⁰⁰ See appendix exhibit 1.3a and 3b for additional results on the achievement information teachers received.

Exhibit 3.2. Percentage of teachers who reported receiving specific types of student achievement information, by treatment status and year

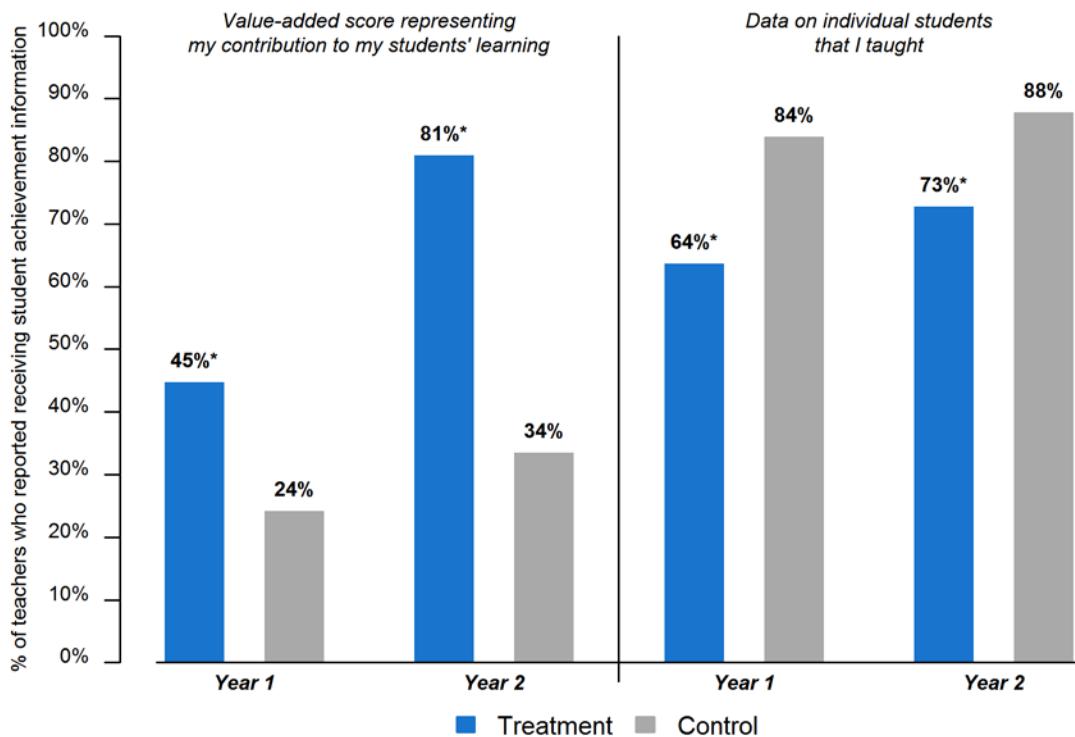


EXHIBIT READS: Of treatment teachers in Year 1, 45 percent reported receiving value-added scores based on the students they taught, compared with 24 percent of control teachers.

NOTES: Year 1 sample size = 63 schools and 519 teachers for the treatment group; 64 schools and 554 teachers for the control group. The analyses were based on a teacher-level regression controlling for random assignment blocks. Year 2 sample size = 63 schools and 492-498 teachers for the treatment group; 63 schools and 521-522 teachers for the control group.

The analyses were based on a teacher-level regression controlling for random assignment blocks.

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

See appendix exhibits I.3a and 3b for separate results for CLASS and FFT districts and for grade K–3 teachers.

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.

SOURCES: Spring 2013 and Spring 2014 Teacher Surveys.

Feedback for Principals

Paralleling the intervention for teachers, the study's intervention for principals was expected to increase the amount and quality of feedback principals received. We tested this theory based on data collected from a principal survey administered just prior to the second VAL-ED wave each year.

In both years, treatment principals reported receiving more feedback with ratings than control principals. Consistent with the design of the intervention, treatment principals reported more instances of oral feedback with ratings than control principals (1.0 versus 0.4 in

Year 1, and 2.0 versus 1.0 in Year 2). (See exhibit 3.3.)¹⁰¹ In addition, as expected, in both years, the average treatment principal reported receiving more oral feedback than did the average control principal (60 versus 41 minutes in Year 1, and 60 versus 33 minutes in Year 2).¹⁰²

Exhibit 3.3. Number of feedback instances and duration of oral feedback that principals reported receiving, by treatment status and year

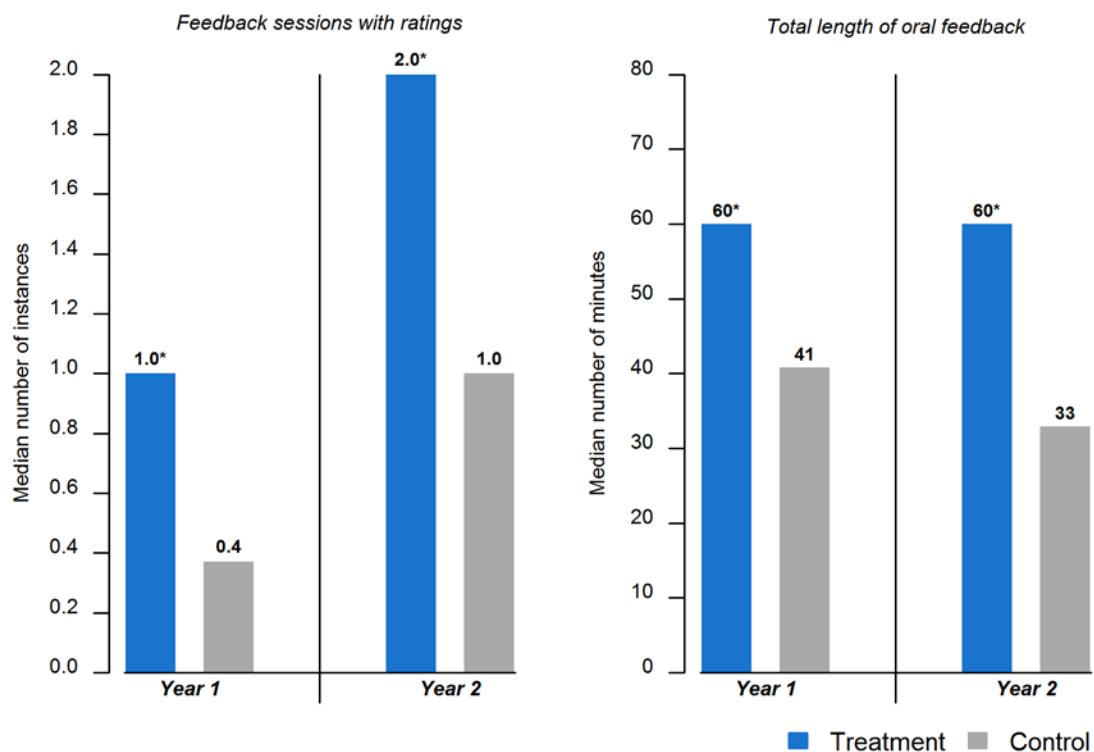


EXHIBIT READS: The average treatment principal in Year 1 reported receiving 1.0 feedback sessions with ratings, compared with 0.4 for control teachers.

NOTES: Year 1 sample size = 61 treatment and 61 control principals. Year 2 sample size = 61 treatment and 59 control principals.

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

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

See appendix exhibits I.4a and 4b for separate results for CLASS and FFT districts.

SOURCES: Spring 2013 and Spring 2014 Principal Surveys.

Impact on Initial Outcomes

According to the study's theory of action, if feedback is frequent and systematic, it may have an impact on educators' "initial outcomes," including their interest in improving along the

¹⁰¹ The principal survey was administered later in the spring in Year 2 than Year 1, permitting principals to include feedback that occurred later in the school year. This may explain why both treatment and control principals reported more instances of feedback in Year 2 than Year 1.

¹⁰² See appendix exhibits I.4a and 4b for analyses conducted separately for districts that used the CLASS and FFT. For CLASS districts, there were no statistically significant treatment-control differences in the number of feedback sessions or the duration of feedback in either year. For FFT districts, treatment principals reported participating in statistically significantly more feedback sessions with ratings than control principals in Year 1, and more hours of feedback in both years, paralleling the overall results.

dimensions on which they received feedback and their perceptions of their own effectiveness. (See exhibit 1.1.) This section presents the results for these hypotheses, based on data from the teacher and principal surveys.

Initial Outcomes for Teachers

Although more treatment than control teachers reported discussing at least one CLASS/FFT-related area with someone who provided feedback, the intervention had little impact on teachers' interest in improving or their participation in professional development in CLASS/FFT-related areas. The intervention was expected to shift the focus of feedback on teacher performance toward areas of classroom practice measured by the CLASS and FFT, and potentially away from areas not measured by the CLASS and FFT. To test this theory, the teacher survey asked teachers which of several areas they discussed with someone who provided feedback on their teaching. The survey item asked about nine areas covering material that almost all teachers might find relevant to improving their instruction (e.g., behavior management or content-specific teaching techniques). Five are areas measured by the CLASS and FFT (behavior management, classroom organization, emotional support for students, instructional dialogue, and student engagement), and four are areas not measured by either the CLASS or FFT (lesson planning, data use, content-specific teaching techniques, and content knowledge).¹⁰³ As expected, in both years, treatment teachers were more likely than control teachers to report discussing at least one of the five CLASS/FFT-related areas with someone who provided them feedback. The intervention had no effect on the percentage of teachers who reported discussing at least one area not related to the CLASS/FFT. (See exhibit 3.4 for Year 2 results and appendix exhibits I.5a, 5b, 6a, and 6b for detailed results for both years.)

Despite the fact that the intervention increased discussion of CLASS/FFT-related areas of practice, teachers did not report greater interest in improving or participating in more professional development in these areas. The study's theory of action posited that feedback on specific areas of practice might lead teachers to seek to improve in those areas, either because their performance had been found to be weaker than desired or because the feedback highlighted the areas as elements of effective teaching.¹⁰⁴ But the theory was not borne out. As exhibit 3.4 shows, treatment teachers were no more likely than control teachers to report interest in improving in at least one area measured by the CLASS and FFT. The intervention also had no impact on teachers' interest in improving in unrelated areas. (See appendix exhibits I.7a, 7b, 8a, and 8b for detailed results for both years.)

The intervention also had no impact on teachers' professional development in CLASS/FFT-related topics in Year 2. But it reduced teachers' participation in professional development covering topics *not* aligned with the CLASS/FFT, perhaps indicating that treatment teachers narrowed the focus of their professional development. There was no impact on teachers'

¹⁰³ The full FFT asks about lesson planning, but that component is not observed during a lesson and so was not included in the feedback provided as part of the study's intervention. See chapter 1.

¹⁰⁴ One might also hypothesize that the intervention would lead teachers who received low scores on their initial CLASS or FFT feedback to want to improve, but it would not have an effect on teachers who received moderate or high scores. We were unable to test this hypothesis because we lacked CLASS and FFT scores for control teachers, other than those collected as outcome measures at the end of Year 2, which was too late to serve as a baseline measure.

participation in professional development on either aligned or nonaligned topics in Year 1 or the summer between Years 1 and 2. (See appendix exhibits I.9a–9c and I.10a–10c for detailed results for Year 1, the summer between Years 1 and 2, and Year 2.)

Exhibit 3.4. Percentage of teachers reporting that they discussed, were interested in improving, and participated in professional development covering at least one area of practice measured by the CLASS or FFT or at least one area not measured, by treatment status, Year 2

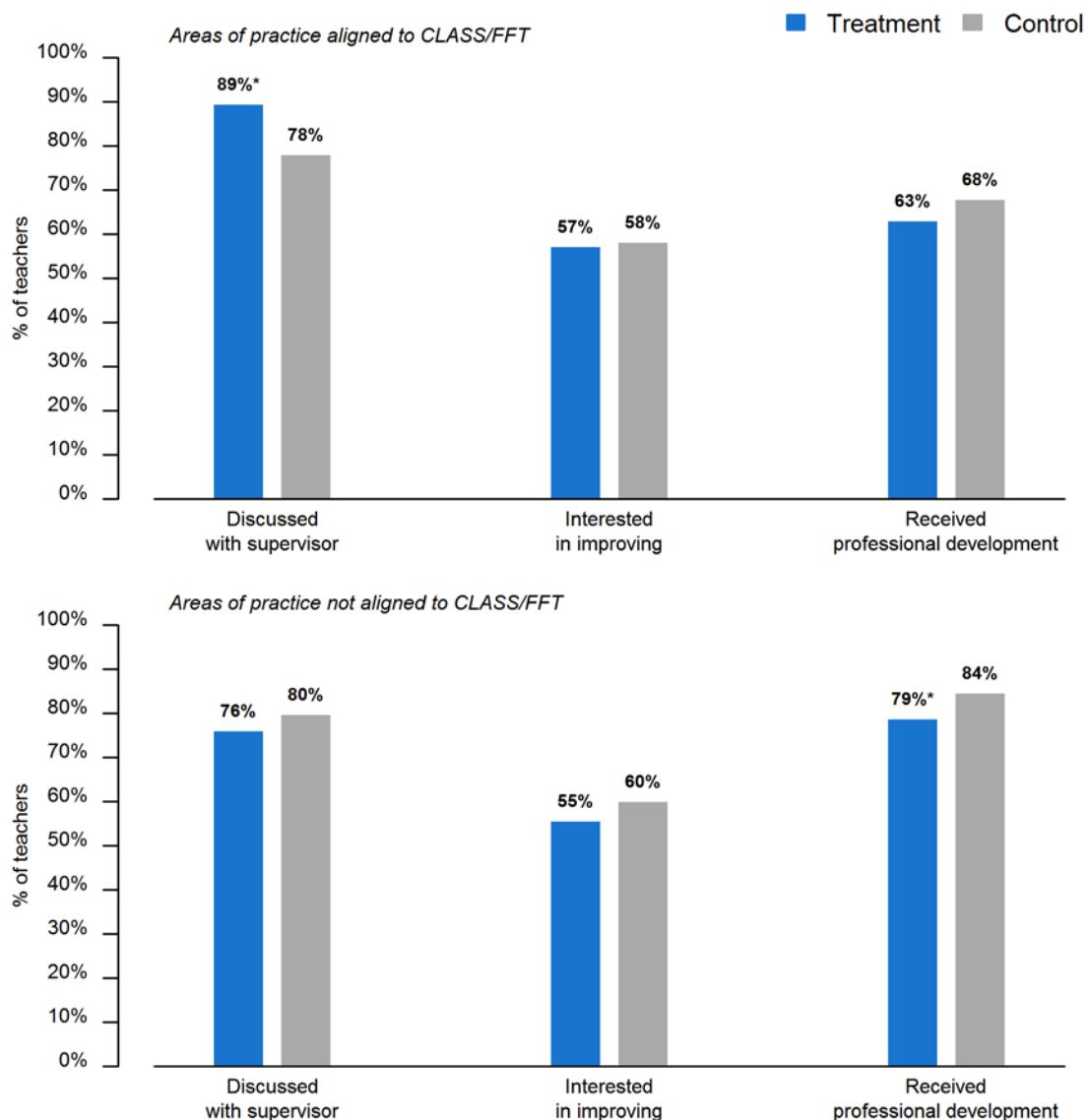


EXHIBIT READS: Of treatment teachers in Year 2, 87 percent reported discussing at least one area of practice measured by the CLASS or FFT with someone who provided them with feedback, compared with 73 percent of control teachers.

NOTES: Sample size = 63 schools and 519 teachers for the treatment group; 64 schools and 554 teachers for the control group. The analyses were based on a teacher-level regression controlling for random assignment blocks and teacher background characteristics.

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

See appendix exhibits I.5b, 6b, 7b, 8b, 9c, and 10c for separate results for CLASS and FFT districts and for grade K–3 teachers.

SOURCE: Spring 2014 Teacher Survey.

The intervention did not lower teachers' perceptions of their effectiveness, and for one aspect of effectiveness (mathematics in Year 1), it raised perceptions.

According to the theory of action underlying the intervention, performance feedback might lower some teachers' perceptions of their effectiveness. In particular, the student growth reports teachers received provided explicit information on teachers' percentile rank in the district in reading/ELA and mathematics. Research on teacher evaluation has observed that traditional forms of evaluation have generally given most or all teachers high ratings (Weisberg et al. 2009). Thus, the value-added information was likely to be more critical than the feedback teachers were accustomed to receiving, and this might lead them to lower their appraisal of their own performance.

To assess whether the intervention lowered teachers' self-appraisal, the survey asked teachers to rate their effectiveness in improving achievement in reading/ELA and mathematics, relative to other teachers in the district.¹⁰⁵ Contrary to what was hypothesized, for reading/ELA, the treatment had no effect on teachers' self-ratings in either year. Treatment and control teachers both rated themselves at the 74th percentile, on average, in Year 1. In Year 2, treatment teachers put themselves at the 72nd percentile and control teachers put themselves at the 73rd percentile. (See exhibit 3.5.) In mathematics, the intervention had an impact in Year 1, but not in the anticipated direction: 78th percentile for treatment teachers versus 75th percentile for control teachers. In Year 2, there was no significant impact on teachers' self-ratings in mathematics (76th percentile for both treatment and control teachers).¹⁰⁶

¹⁰⁵ The survey asked teachers to assess their performance using a set of six performance categories: very poor (bottom 5 percent); poor (6th to 25th percentile); fair (26th to 50th percentile); good (51st to 75th percentile); very good (76th to 95th percentile); and exceptional (top 5 percent). To compute the average percentile for treatment and control teachers, we replaced each performance category with the midpoint of the percentile range for that category: 3, 15.5, 38, 63, 85.5, and 98. (See appendix I for details about the analyses.)

¹⁰⁶ See appendix exhibit I.11 for analyses conducted separately for districts using the CLASS and FFT. In CLASS districts, treatment teachers had higher self-ratings than control teachers in both reading/ELA and mathematics in Year 1; there were no statistically significant differences in Year 2. In FFT districts, treatment teachers had lower self-ratings than control teachers in reading/ELA in Year 1; otherwise, there were no statistically significant differences.

Exhibit 3.5. Teachers' self-ratings of their effectiveness in boosting students' reading/ELA and mathematics achievement, by treatment status and year

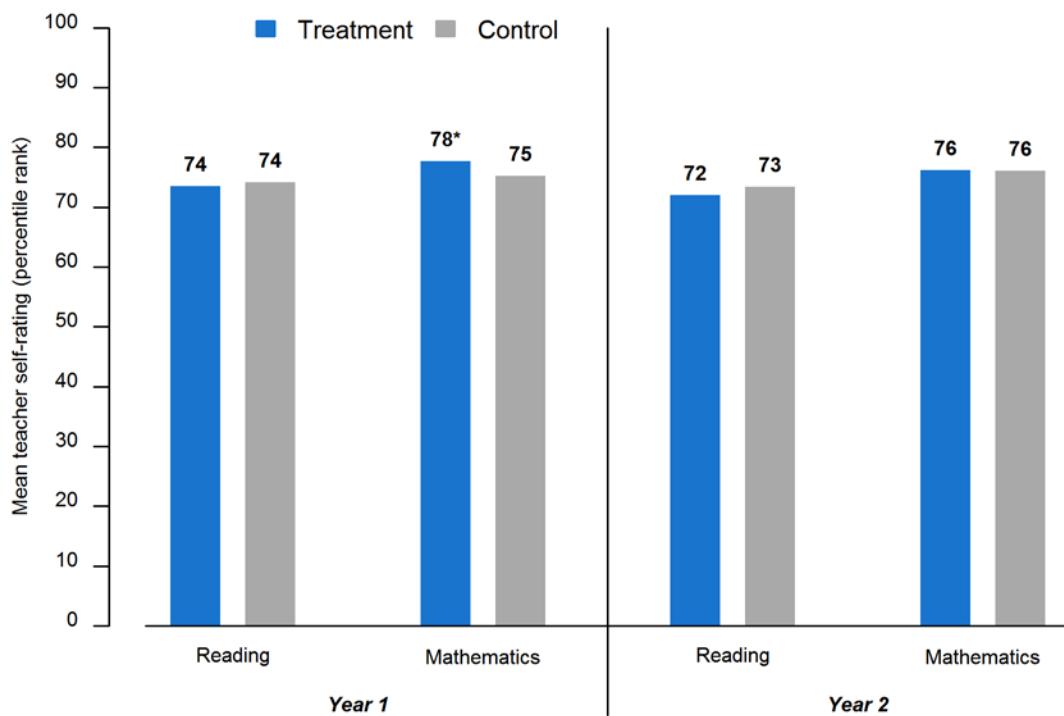


EXHIBIT READS: The average self-rating for treatment teachers in Year 1 was at the 74th percentile, as was the average rating for control teachers.

NOTES: Year 1 sample size = 63 schools and 425–428 teachers for the treatment group; 64 schools and 437–441 teachers for the control group. Year 2 sample size = 63 schools and 398–401 teachers for the treatment group; 63 schools and 396–414 teachers for the control group.

The analyses were based on a teacher-level regression controlling for random assignment blocks and teacher background characteristics.

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

See appendix exhibits I.11 for separate results for CLASS and FFT districts.

SOURCES: Spring 2013 and Spring 2014 Teacher Surveys.

Rather than lowering teachers' self-appraisal on average, perhaps receiving performance information might cause the teachers' self-ratings to become more aligned with their true performance, as measured by their value-added score. To test this hypothesis, we examined whether the association between teachers' value-added score (in percentile units) and their self-rating (also in percentile units) was stronger among treatment than control teachers. This hypothesis was not supported either in Year 1 or 2. (See appendix exhibit I.12a.) The association between teachers' value-added and their self-perceived effectiveness was modest in both treatment and control conditions, and the association was no stronger for treatment than for control teachers.¹⁰⁷

¹⁰⁷ See Exhibits I.12b and 12c for plots of the relationship between teachers' value-added and self-perceived effectiveness.

Initial Outcomes for Principals

The intervention did not affect the percentage of principals discussing areas related to VAL-ED with their supervisors, their interest in improving, or their participation in professional development covering these areas. The principal survey asked the principals whether they discussed various areas with their supervisors. The survey item asked about seven areas, all covering material that all principals might find relevant to improving their leadership (e.g., advising teachers on ways to improve their instruction, or making personnel/human resource decisions). The survey asked about four areas measured by the VAL-ED core components (identifying, implementing, or monitoring the use of challenging curriculum; advising teachers on ways to improve their instruction; using data to make decisions; and parent/community issues). It also asked about three areas not measured by VAL-ED (making personnel/human resources decisions; managing nonpersonnel administrative issues; and student behavior/discipline). Based on the theory of action, we expected the intervention to lead principals to discuss areas related to the VAL-ED with their supervisors. But contrary to expectations, in both years, treatment principals were no more likely than control principals to report discussing at least one area related to the VAL-ED with their supervisor. The intervention also had no impact on the percentage of principals who reported discussing at least one unrelated area. (See exhibit 3.6 for Year 2 results and appendix exhibits I.13a, 13b, 14a, and 14b for detailed results for both years.)¹⁰⁸

The study's theory of action also suggested that the intervention might increase treatment principals' interest in improving in areas related to the VAL-ED. This hypothesis was also not supported. In both years, treatment principals were no more likely than control principals to report wanting to improve in at least one area related to the VAL-ED, and no less likely to report wanting to improve in at least one unrelated area. (See exhibit 3.6 for results for Year 2 and appendix exhibits I.15a, 15b, 16a, and 16b for detailed results for both years.)

Similarly, the intervention did not lead treatment principals to participate in more professional development in areas related to VAL-ED. Paralleling the theory of action for teachers, the theory assumed that the feedback might lead principals to seek professional development on topics related to the VAL-ED either to strengthen areas of identified weakness or to learn more about areas emphasized in the VAL-ED. However, the results showed no impact of the intervention on principals' professional development. (See exhibit 3.6 for results for Year 2 and appendix exhibits I.17a, 17b, 17c, 18a, 18b, and 18c for detailed results for Year 1, the summer between Years 1 and 2, and Year 2.)

¹⁰⁸ In districts that used CLASS, the intervention had a negative impact on discussing topics not measured by the VAL-ED in both Years 1 and 2. In districts that used the FFT, the intervention had a positive impact on discussing topics measured by the VAL-ED in Year 1 but not Year 2.

Exhibit 3.6. Percentage of principals reporting that they discussed, were interested in improving, and participated in professional development covering at least one area of practice measured by the VAL-ED, by treatment status, Year 2

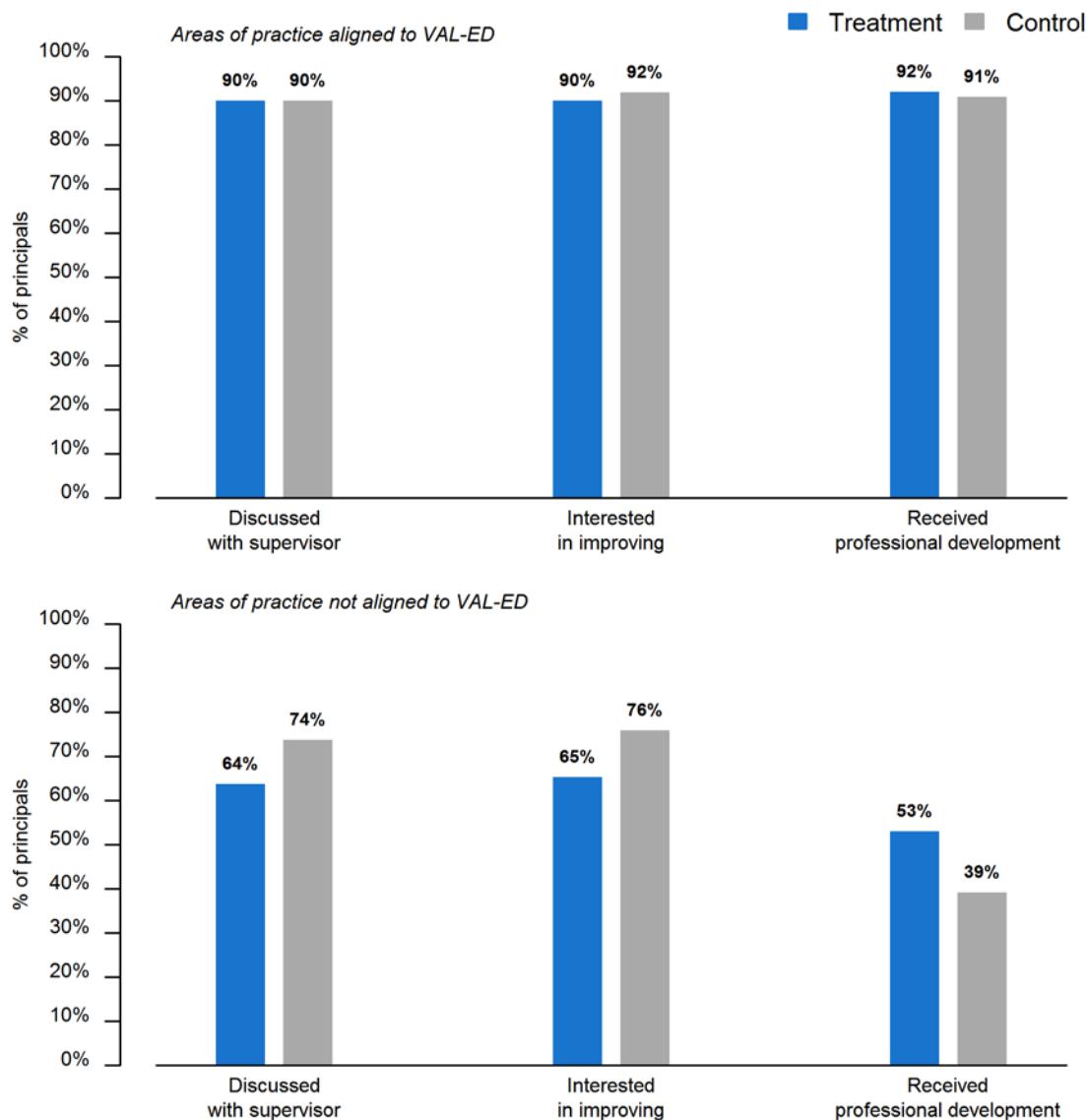


EXHIBIT READS: Of treatment principals in Year 1, 92 percent reported discussing at least one area of practice measured by the VAL-ED with a supervisor, compared with 88 percent of control principals.

NOTES: Sample size = 63 treatment principals and 63 control principals.

The analyses were based on a principal-level regression controlling for random assignment blocks and principal background characteristics.

None of the differences between the treatment and the control groups are statistically significant at the .05 level (two-tailed).

See appendix exhibits I.13b, 14b, 15b, 16b, 17c, and 18c for separate results for CLASS and FFT districts.

SOURCE: Spring 2014 Principal Survey.

The intervention did not change principals' perceptions of their effectiveness. We tested a hypothesis for principals similar to the one for teachers discussed above—that the VAL-ED might provide more critical feedback than principals were accustomed to,¹⁰⁹ and that this might change their self-perception. To test this, the Year 2 principal survey asked principals to rate their effectiveness relative to other principals in two domains: providing instructional leadership and other forms of leadership.¹¹⁰

The results show no statistically significant difference in self-ratings for principals in the treatment and control conditions, in either instructional leadership or other forms of leadership. Treatment principals' average self-rating for instructional leadership in Year 2 was the 76th percentile, compared to the 73rd percentile for control principals. Similarly, treatment principals put themselves in the 80th percentile for other forms of leadership, compared to the 79th percentile for control principals. (See exhibit 3.7.)

¹⁰⁹ Unlike the value-added reports teachers received, the VAL-ED reports did not provide principals' percentile rank in comparison to others in their district, although it did show how principals scored in relation to the VAL-ED national norming sample. This may have attenuated any impact of providing the VAL-ED on principal perceptions.

¹¹⁰ This item was not included on the Year 1 principal survey.

Exhibit 3.7. Principals' self-rating of their effectiveness in instructional leadership and other forms of leadership, by treatment status, Year 2

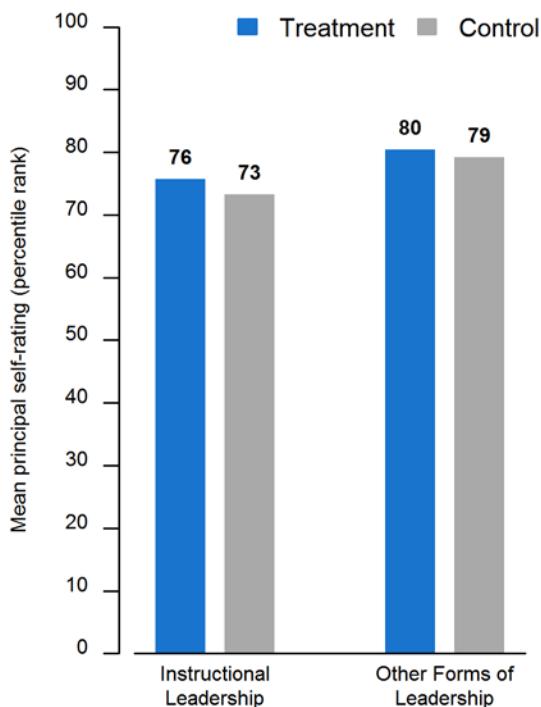


EXHIBIT READS: The average self-rating for treatment principals in the area of Instructional Leadership in Year 2 was at the 76th percentile, compared with the 73rd percentile for control principals.

NOTES: Sample size for instructional leadership = 61 treatment and 60 control principals. Sample size for other forms of leadership = 61 treatment and 59 control principals.

The analyses were based on a principal-level regression controlling for random assignment blocks and principal background characteristics.

None of the differences between the treatment and the control groups are statistically significant at the .05 level (two-tailed).

See appendix exhibit I.19 for separate results for CLASS and FFT districts.

SOURCE: Spring 2014 Principal Survey.

Impact on Classroom Practice, Principal Leadership, and Student Achievement

The primary assumption underlying the performance feedback was that it would improve teacher classroom practice, principal leadership, and ultimately student achievement. The theory of action assumed that impacts on these outcomes could occur through at least two mechanisms. First, positive feedback on their performance could lead more-effective teachers and principals to remain in their schools, while negative feedback could lead less-effective staff to leave, opening their positions to be filled by more-effective staff. Second, feedback might improve the practice of teachers and principals who stayed. As described at the start of the chapter, the analyses focus on all teachers and principals present in the study schools in the spring of Years 1 and 2, and thus the sample includes some educators who stayed, as well as some who were new to their schools. Any impacts observed thus reflect a mix of the two processes hypothesized to lead to improvement in educator and student outcomes.

To assess whether any estimated impacts on classroom practice, leadership, or achievement might have been caused by differences between treatment and control schools in overall teacher or principal mobility, we estimated the impact of the intervention on teacher and principal exits between Year 1 and 2.¹¹¹ Exhibit A.9 shows the results of these analyses. The results indicate that there was no statistically significant impact, although the impact on teacher exits was negative and close to significant ($p = 0.053$). This suggests that overall differences in mobility are not likely to explain any observed impacts on outcomes. It is possible, however, that the intervention may have had different effects on the mobility of more- and less-effective teachers, which could have affected observed outcomes.¹¹²

The following sections report results for classroom practice, principal leadership, and student achievement. For classroom practice, we gathered data only in Year 2, because it required a resource-intensive process of video-recording and coding. For both leadership and achievement, we examined outcomes in both years.¹¹³

Impact on Classroom Practice

To provide a common measure to use in assessing the impact of the intervention on teacher classroom practice, we video-recorded one lesson for each treatment and control teacher in the spring of Year 2 and a second lesson for a random sample of half the teachers. Each lesson was coded by trained observers on the study research team, using *both* the CLASS and the FFT instruments. Both instruments were used as an outcome measure in all study districts in order to assess whether the feedback had an impact on the specific practices measured by the instrument on which the feedback was based, as well as on other practices not as specifically targeted. (See appendix B for more information on the video-recording and coding procedures.)

The intervention had a positive impact on teacher classroom practice based on video-recorded lessons coded using the CLASS, but not on practice coded using the FFT. On average, treatment teachers received a score of 4.50 on the CLASS, based on the 7-

¹¹¹ We also conducted parallel analyses examining the impact on student exits between Year 1 and 2 and found no impact. See appendix exhibit A.9.

¹¹² We tested the baseline equivalence of treatment and control teachers and principals in the Year 2 analysis sample for classroom practice and the Year 1 and 2 samples for principal leadership. (Results are shown in appendix exhibits J.1–5.) There were three statistically significant differences between the treatment and control groups in measured baseline characteristics. The differences all pertain to the teacher experience background variables in the Year 1 and 2 principal leadership impact samples. (See appendix exhibits J.2 and 3). To take these differences into account, the experience measures were included as covariates in the impact models. We also tested the baseline equivalence of treatment and control students in the Year 1 and 2 analysis samples for student achievement. (Results are shown in exhibits J.6–13). There were no statistically significant differences in measured characteristics.

¹¹³ We conducted supplementary analyses to test the sensitivity of the results on the impact of the intervention on classroom practice, principal leadership, and student achievement to details of the analytic model. The basic pattern of the results was unchanged. (See appendix H for a discussion of the models estimated and appendix exhibits J.18, 19, 24, 30, 31, 32, and 33, for the results.) For the classroom practice analysis, one CLASS district experienced lower video response rates in the treatment than the control. When that district was excluded from the analysis, the effect of the intervention on the CLASS practice measure was not significant across the seven districts, but it was significant for the CLASS districts. (See appendix exhibit J.19.) For the student achievement analysis, the Year 2 mathematic impact that was not statistically significant in the main analysis was significant in a sensitivity analysis including additional covariates. (See appendix exhibit J.32.)

point CLASS scale, compared with 4.39 for control teachers.¹¹⁴ (See exhibit 3.8.) The 0.11-point difference corresponds to an effect size of 0.17 and an improvement index of 7 percentile points, implying that the percentile rank of the average control teacher would increase from the 50th percentile to the 57th percentile if the teacher received the intervention. There was no statistically significant difference between the treatment and control teachers when classroom practice was measured based on video-recorded lessons coded using the FFT.¹¹⁵

While the results indicate that the intervention had a positive impact on classroom practice, it is not clear how this impact was generated. The study's theory of action hypothesized that the intervention would operate through a set of initial outcomes, including discussing CLASS/FFT topics with staff providing feedback, an interest in improving in CLASS/FFT areas, participating in professional development in these areas, and a change in self-perceived effectiveness. But with the exception of the first of these, the results indicate that the intervention did not have the anticipated impacts on initial outcomes. Perhaps the feedback had an effect on initial outcomes that were not captured on the study's teacher survey; for example, perhaps it led teachers to increase the amount of class time spent in instruction. Or, perhaps the feedback itself provided teachers insight into their teaching, which led directly to improved practice. (See Rowan and Raudenbush 2016, for a discussion of this hypothesis.) We lack data to test this theory, however.

¹¹⁴ We also conducted exploratory analyses of the impact of the intervention on the four CLASS domain scores and the two FFT domain scores. (See appendix exhibits J.15 and 16.)

¹¹⁵ We examined whether the impact of the intervention on classroom practice differed for probationary and nonprobationary teachers, for teachers of elementary and middle schools, and for teachers with lower and higher baseline value-added scores. We found only one statistically significant differential effect among these teacher groups: the impact on classroom practice as measured by the CLASS was larger for teachers with lower prior valued-added scores than for teachers with higher prior scores. (See appendix exhibit J.34.)

Exhibit 3.8. Average CLASS and FFT scores, based on coding of video-recorded lessons by study team, by treatment status, Year 2

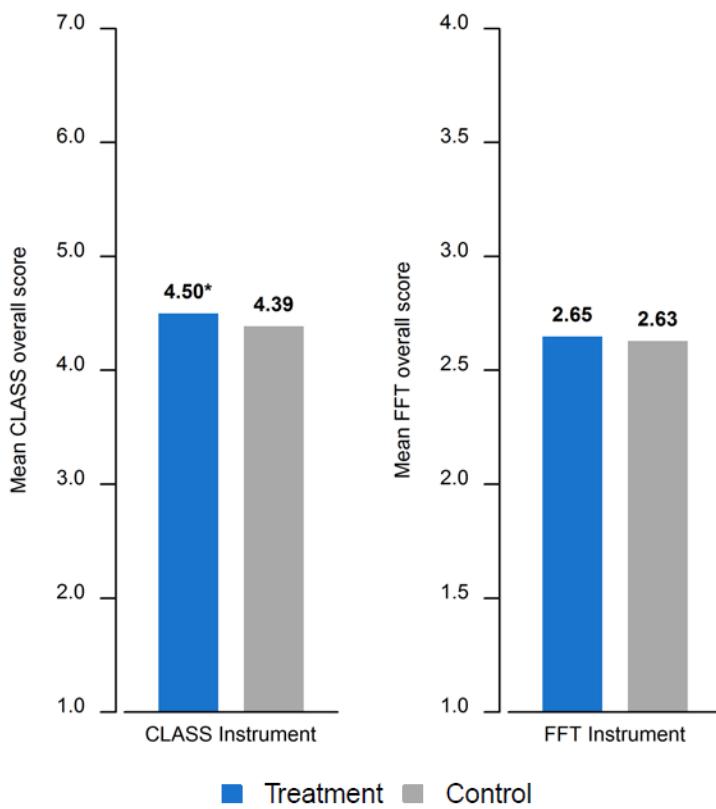


EXHIBIT READS: The average CLASS overall score was 4.50 for treatment teachers, compared with 4.39 for control teachers.

NOTES: Sample size = 63 schools, 434 teachers, and 668 videos for the treatment group; 63 schools, 517 teachers, and 793 videos for the control group.

The analyses were based on a three-level regression (lessons within teachers within schools) controlling for random assignment blocks and teacher background characteristics.

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

See appendix exhibits J.14 for additional detail, J.15 and 16 for results for CLASS and FFT domain scores, and J.17 for results by study district.

SOURCE: Spring 2014 Classroom Videos.

In addition to estimating the average impact of the intervention on classroom practice, we examined the consistency of the impact across the eight districts, to assess potential variation due to district context or policy. We found statistically significant variation across districts in the impact of the intervention on classroom practice as measured by the CLASS, but not as measured by the FFT. (See appendix exhibit J.17.) The impact on practice as measured by the CLASS ranged from -0.13 to 0.56 across the eight study districts, with three positive and statistically significant (districts 2, 3, and 4). The impact on practice as measured by the FFT ranged from -0.07 to 0.07 across the eight study districts, with none statistically significant.

Using the video-recorded lessons, we also estimated the impact separately for the four districts that used the CLASS for feedback and the four that used the FFT. If the intervention operates as intended, it should lead teachers to improve in the areas of practice explicitly targeted (e.g., CLASS scores in districts that used the CLASS for feedback, and FFT scores in districts that used the FFT for feedback). This is the most proximal outcome in the theory of action.

In addition to examining the impact on the measure used for feedback, we tested the impact on the FFT in CLASS districts and vice versa. Although the CLASS and FFT generally tap similar dimensions, the two instruments give the dimensions different degrees of emphasis and define them somewhat differently. Using both the CLASS and the FFT as outcome measures provides evidence about whether the impact of the intervention extends to related areas of practice, in addition to the specific dimensions targeted.

We found a positive impact on CLASS scores in the four CLASS districts, but not in the four FFT districts. (See exhibit 3.9.) On average, across the four CLASS districts, treatment teachers received a CLASS score of 4.64, compared with 4.32 for control teachers. The 0.31-point difference corresponds to an effect size of 0.46 and an improvement index of 18 percentile points, meaning the percentile rank of an average control teacher would increase from the 50th to the 68th percentile if the teacher received the intervention.^{116,117} The difference between the impact on CLASS scores in the CLASS districts and the impact in the FFT districts was statistically significant. (See appendix exhibit J.14.) There was no impact on the FFT in either CLASS or FFT districts.

Because study districts chose to use the CLASS or the FFT as part of the intervention, we cannot draw definitive conclusions about why an impact on classroom practice was found in CLASS but not in FFT districts. Aspects of the CLASS and FFT measures, reports, and feedback sessions could have influenced the results. For example, the CLASS measure used a 7-point rating scale, while the FFT used a 4-point scale, which could have influenced the way performance information was communicated to teachers. With respect to the reports, most CLASS reports identified at least one dimension of classroom practice to improve and illustrated it with an example from the observation; fewer FFT reports did so. With respect to activities during the feedback sessions, teachers were more likely to watch video clips illustrating strong performance during CLASS than during FFT feedback sessions, which may account for the pattern of results in CLASS and FFT districts.

The results could also be due to features of district policy or demographic context, or to features of the CLASS and FFT feedback systems. For example, because teachers and principals in treatment schools were expected to receive the feedback ordinarily included as part of their districts' evaluation systems as well as the study feedback, variation in these systems could potentially have affected the impact of the study feedback. As shown in exhibits 1.4 and 1.5, some districts provided more observer training than others, or included more frequent feedback.

¹¹⁶ The value of 0.31 points differs from the observed treatment-control difference due to rounding.

¹¹⁷ For comparison, in a recent randomized study of My Teaching Partner (Allen et al. 2011), which provided about nine rounds of structured feedback using the CLASS over a single year, the intervention had an impact of about 0.74 standard deviations on instruction, as measured using a composite of five dimensions of the CLASS.

Exhibit 3.9. Average CLASS and FFT scores in CLASS districts and FFT districts, based on coding of video-recorded lessons by study team, by treatment status, Year 2

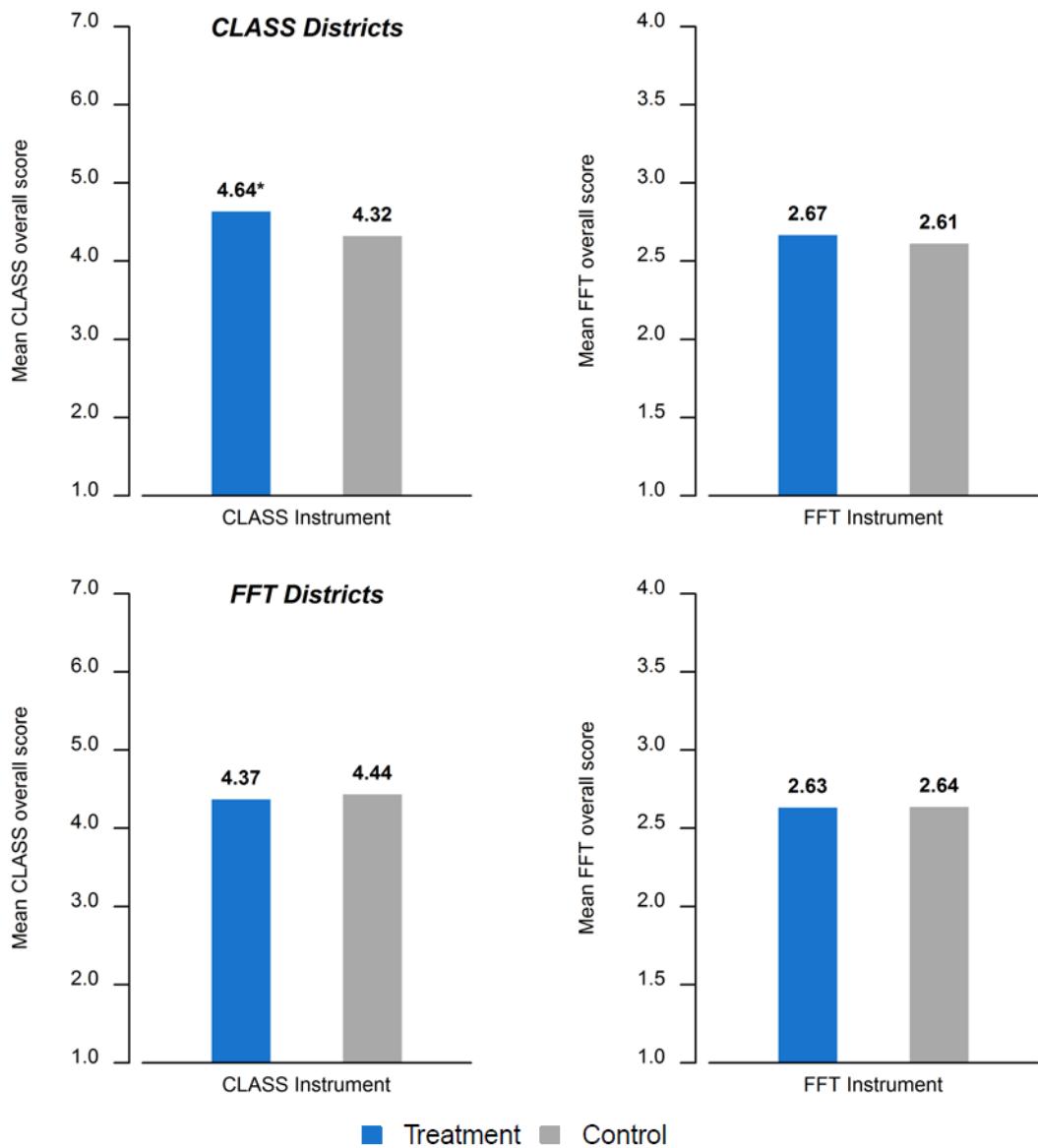


EXHIBIT READS: In CLASS districts, the average CLASS overall score in Year 2 was 4.64 for treatment teachers, compared with 4.32 for control teachers.

NOTES: Sample size for CLASS districts = 63 schools, 238 teachers, and 360 videos for the treatment group; 63 schools, 306 teachers, and 462 videos for the control group. Sample size for FFT districts = 63 schools, 211 teachers, and 308 videos for the treatment group; 63 schools, 232 teachers, and 331 videos for the control group.

The analyses were based on a three-level regression (lessons within teachers within schools) controlling for random assignment blocks and teacher background characteristics.

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

See appendix exhibits J.14 for additional detail and J.15 and 16 for results for CLASS and FFT domain scores.

SOURCE: Spring 2014 Classroom Videos.

Impact on Principal Leadership

The main goal of the VAL-ED feedback for principals was to improve their leadership skills. By giving increased attention to teaching and learning (the focus of the VAL-ED) and by spending time observing and discussing classroom practice (the focus of the CLASS/FFT measures), the principal may come to be perceived by teachers as a trusted instructional leader. To assess this outcome, we relied on two measures using items on the spring Year 1 and Year 2 teacher survey, based on scales developed by the Chicago Consortium on School Research (CCSR 2012): instructional leadership and teacher-principal trust. The instructional leadership scale measures teachers' perceptions of their principal as an instructional leader, for example whether the principal sets high standards for teaching and learning, actively monitors the quality of teaching, and has clear expectations about instructional goals (Sebastian and Allensworth 2012). These items are similar to four VAL-ED core components: High standards for teaching, Rigorous curriculum, Quality instruction, and Performance accountability. Teacher-principal trust measures the extent to which teachers feel their principal has established trusting relations with them, for example by taking an interest in them as professionals, being responsive to their input, and placing the needs of children ahead of personal interests. These items are similar to one VAL-ED core component: Culture of learning and professional behavior. The VAL-ED core component Connection to external communities is not reflected in either of the scales. (See appendix B for details about the two scales.)

We chose items based on CCSR scales rather than items from the VAL-ED to assess leadership because of treatment teacher experience with the VAL-ED. By the time of the Year 1 spring survey, most treatment teachers had already completed the VAL-ED survey twice, and by the Year 2 spring surveys, a large majority of treatment teachers had already completed the VAL-ED four times, making it likely that they would respond to the survey with a disposition or framework different from that used by control teachers, who had never before completed a VAL-ED survey. Even though the items based on the CCSR scales differ from the VAL-ED, treatment teachers' responses may still have been affected by their experience with VAL-ED in ways that could have contributed to the estimated impacts.

The intervention had a positive impact on teacher-principal trust in Year 1, and on both instructional leadership and teacher-principal trust in Year 2. In Year 1, treatment principals on average received a score of 3.18 on the 5-point teacher-principal trust scale, compared with 2.96 for control principals. (See exhibit 3.10.) The 0.22-point difference corresponds to an effect size of 0.25 and an improvement index of 10 percentile points, implying that the trust score for the average control principal would increase from the 50th percentile to the 60th percentile if the school received the intervention. In Year 2, there were positive impacts on both instructional leadership (0.14 points) and teacher-principal trust (0.15 points).¹¹⁸ Although there were statistically significant impacts on both leadership measures in Year 2, and only one in Year 1, the magnitudes of the impacts were similar in the two years, and thus there is little evidence of an increase in effects over the two years.

¹¹⁸ We examined whether the impact of the intervention on principal leadership differed for principals in elementary and middle schools and found no statistically significant differential effects. (See appendix exhibit J.35.)

Exhibit 3.10. Average rating of principal instructional leadership and teacher-principal trust, by treatment status and year

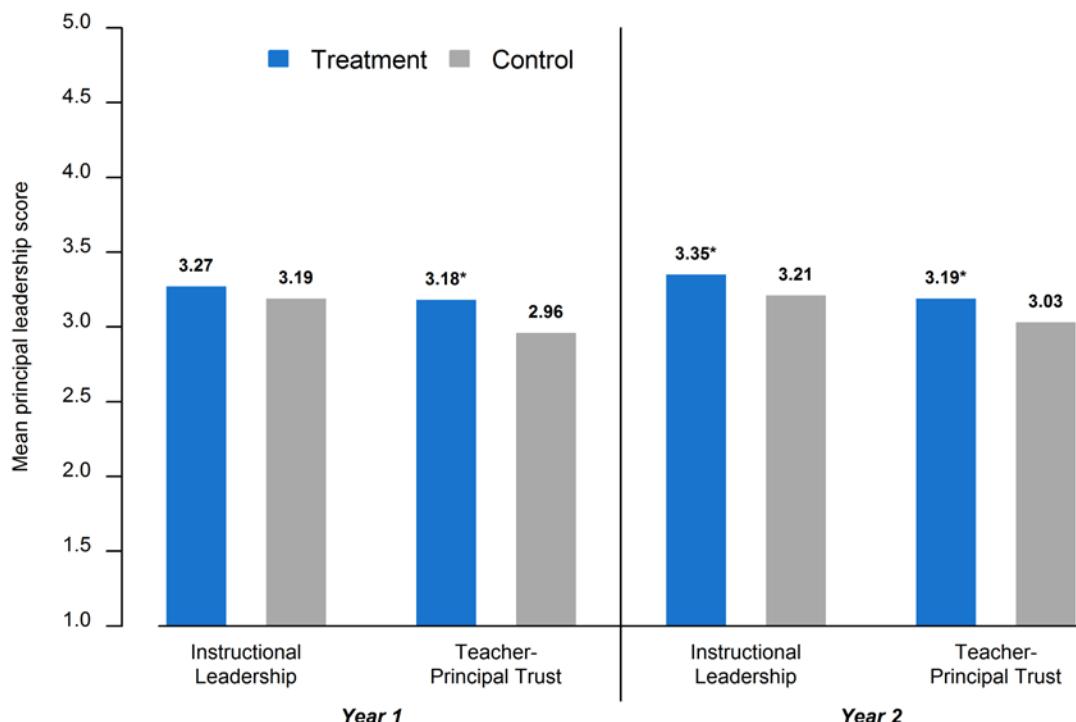


EXHIBIT READS: The average rating of principals' instructional leadership in treatment schools in Year 1 was 3.27, compared to 3.19 for principals in control schools.

NOTES: Year 1 sample size = 63 principals and 524 or 525 teachers for the treatment group; 64 principals and 557 teachers for the control group. Year 2 sample size = 63 principals and 499 teachers for the treatment group; 63 principals and 522 or 523 teachers for the control group.

The analyses were based on a two-level regression (teachers nested in schools) controlling for random assignment blocks and teacher background characteristics.

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

See appendix exhibits J.20 for additional details and J.22 and J.23 for results by district.

SOURCES: Spring 2013 and 2014 Teacher Surveys.

As we did when looking at the impact of the intervention on classroom practice, we examined variation in the impact on leadership across the eight study districts to examine the consistency of the impact. We found statistically significant variation in impact across districts for both instructional leadership and teacher-principal trust in Year 2, but not in Year 1. (See appendix exhibit J.22 and J.23.) The Year 2 impact on instructional leadership ranged from -0.36 to 0.43 across the eight districts, with two positive and statistically significant (districts 2 and 8). The Year 2 impact on teacher-principal trust ranged from -0.65 to 0.53 across the eight districts, with three statistically significant (negative in district 1, positive in districts 2 and 8). This provides evidence that features of district context or policy may have played a role in the effectiveness of the intervention in improving principal leadership.

Principals in all eight study districts received feedback based on the VAL-ED, so we did not anticipate different effects on leadership in CLASS and FFT districts. Because we found positive effects on classroom practice as measured by the CLASS in CLASS districts but not in FFT districts, we wondered if similar effects might have occurred for leadership, perhaps reflecting district differences in the implementation of the intervention, or differences in district context.¹¹⁹ Thus, we conducted an exploratory analysis of impact separately in CLASS and FFT districts. We found that in CLASS districts, the intervention did not have a statistically significant impact on either of the two leadership measures in either year. (See exhibit 3.11.) In FFT districts, the intervention had a positive impact on teacher-principal trust in Year 1 and on both measures of leadership in Year 2, paralleling the overall impact results for leadership.¹²⁰

¹¹⁹ It is also possible that teachers' ratings of their principal as a leader might have been influenced by whether the principal used the CLASS or FFT observation rubric to provide instructional feedback.

¹²⁰ Because the impacts on leadership in CLASS and FFT districts are estimated with error, the apparent differences in results could be due to chance. An exploratory test of the differential impact of the intervention in the two sets of districts yielded a statistically significant result for only one outcome (teacher-principal trust in Year 2). (See appendix exhibit J.21.) Thus, there is little evidence of a systematic difference in impact in CLASS and FFT districts.

Exhibit 3.11. Average rating of principal instructional leadership and teacher-principal trust in CLASS districts and FFT districts, by treatment status and year

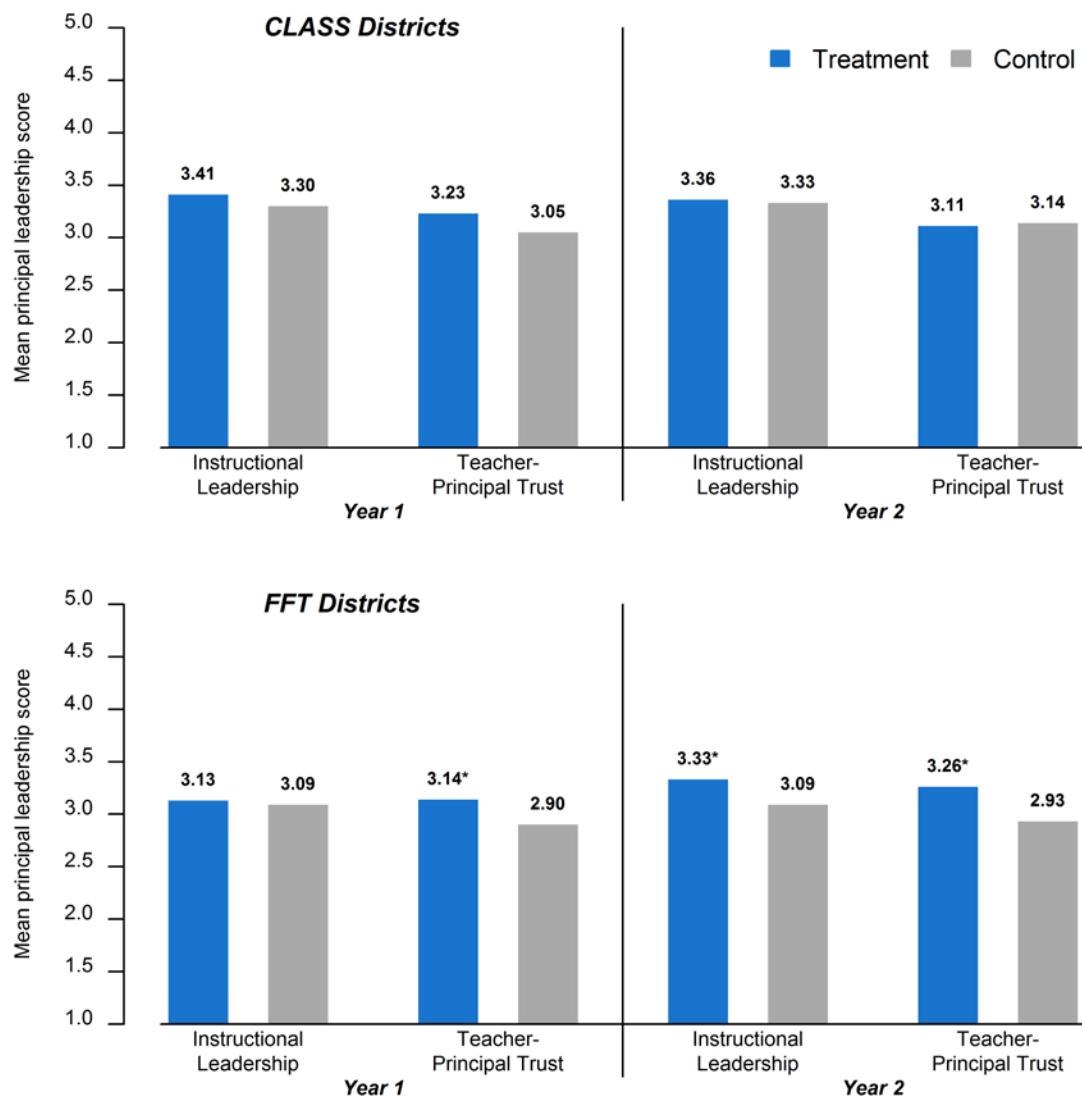


EXHIBIT READS: In CLASS districts, the average rating of principals' instructional leadership in treatment schools in Year 1 was 3.41, compared to 3.30 for principals in control schools.

NOTES: Year 1 sample size for CLASS districts = 31 principals and 307 teachers for the treatment group; 32 principals and 328 teachers for the control group. Year 1 sample size for FFT districts = 32 principals and 217 or 218 teachers for the treatment group; 32 principals and 229 teachers for the control group. Year 2 sample size for CLASS districts = 31 principals and 301 teachers for the treatment group; 32 principals and 312 or 313 teachers for the control group. Year 2 sample size for FFT districts = 32 principals and 198 teachers for the treatment group; 31 principals and 210 teachers for the control group.

The analyses were based on a two-level regression (teachers nested in schools) controlling for random assignment blocks and teacher background characteristics.

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

See appendix exhibit J.21 for additional details.

SOURCES: Spring 2013 and Spring 2014 Teacher Surveys.

Impact on Student Achievement

The ultimate goal of the intervention was to boost students' achievement in reading/ELA and mathematics. We examined the impact on achievement by comparing students' scores on the state achievement test for all students enrolled in treatment and control teachers' classes in the spring of Years 1 and 2. The Year 1 estimates controlled for student achievement in the spring of the year before the intervention was implemented (i.e., the baseline year), and thus the estimates represent the effect of the first year of implementation of the intervention. The Year 2 estimates also controlled for student achievement from the baseline year, and thus they represent the cumulative impact of the intervention over two years.¹²¹ (See appendix H for more detail on the analysis.)

The intervention had a positive impact on students' mathematics achievement in Year 1, and a cumulative impact in Year 2 that was similar in magnitude but not statistically significant ($p = 0.055$). The intervention did not have an impact on students' reading/ELA achievement in either year. In Year 1, students in treatment schools scored at the 51.8th percentile in mathematics in their district, compared to the 49.7th percentile for control students. (See exhibit 3.12.) The 2.1-point difference corresponds to an effect size of 0.05, or about one month of learning.¹²² In Year 2, students in treatment schools scored at the 51.2nd percentile on average, compared to the 48.9th percentile for control students, a 2.3-point difference, similar in magnitude to the impact in Year 1, but not statistically significant ($p = 0.055$).¹²³ The impacts for reading/ELA (0.4 percentile points in Year 1 and 1.0 in Year 2) were smaller than the impacts for mathematics and were not statistically significant.¹²⁴ There is no evidence that the cumulative impact on achievement increased from the first to the second year of implementation.¹²⁵

¹²¹ The Year 2 estimates are based on students in grades 4–8 in the spring of Year 2, including students who were in the study schools in both years, as well as those who entered in Year 2. (See appendix exhibits A.11 and 12 for a description of student entries and exits.) The treatment and control students included in the achievement impact analyses did not differ in demographic characteristics or prior achievement. (See appendix exhibits J.6–13.)

¹²² According to Hill et al. (2008), the average annual gain in mathematics is about 0.42 for students in grades 4–8. Thus, an impact of 0.05 standard deviations translates into about $0.05/0.42 = 0.11$ of a year's achievement gain. Assuming a 36-week school year, this implies that the impact corresponds to four weeks of learning.

¹²³ The Year 2 impact model controls for students' achievement in the baseline year, two years prior to the outcome. Therefore, the variance explained by covariates is somewhat lower in the Year 2 than the Year 1 model, reducing the precision of the Year 2 impact estimates. The impact model for reading/ELA controlled for prior reading/ELA achievement, and the model for mathematics controlled for prior mathematics achievement. As a sensitivity analysis, we estimated the impact models using prior achievement in *both* reading/ELA and mathematics as controls. The estimates are similar, but the impact on mathematics in Year 2 is statistically significant. (See appendix exhibits J.32 and 33.)

¹²⁴ We examined whether the impact of the intervention on achievement differed for students of probationary and nonprobationary teachers; for students in elementary and middle schools; and for students of teachers with lower and higher baseline value-added scores. We found no differential impact. (See appendix exhibit J.36.)

¹²⁵ We are not sure why providing a second year of feedback did not lead to an increase in the cumulative impact. Perhaps teachers took advantage of easy-to-implement recommendations in the first year and would have needed more support in the second year to make additional progress. Or perhaps teachers gave the feedback less attention in the second year because it was no longer novel. Or perhaps the improvements teachers made in the first year were not sustained over the summer. We lack evidence to provide tests of these hypotheses.

Exhibit 3.12. Average reading/ELA and mathematics achievement, by treatment status and year

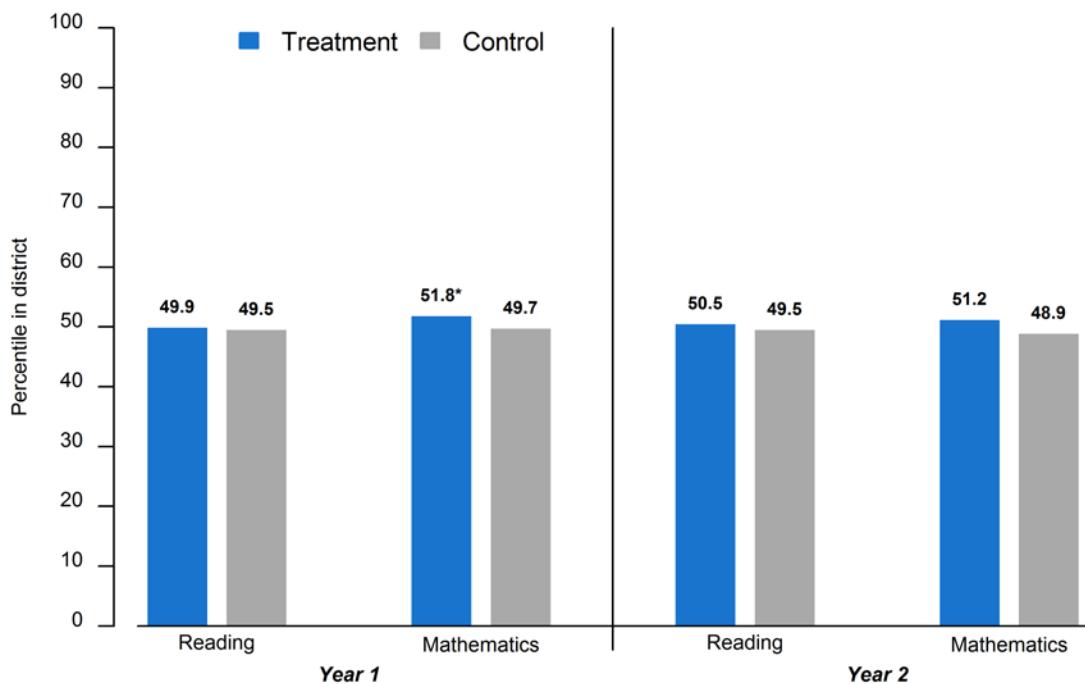


EXHIBIT READS: In Year 1, students in treatment schools received an average reading/ELA score at the 49.9th percentile in their district, compared to the 49.5th percentile for students in control schools.

NOTES: Sample size for Year 1 reading/ELA = 63 schools, 384 teachers, and 13,134 students for the treatment group; 64 schools, 421 teachers, and 15,358 students for the control group. Sample size for Year 1 mathematics = 63 schools, 411 teachers, and 13,967 students for the treatment group; 64 schools, 439 teachers, and 15,907 students for the control group. Sample size for Year 2 reading/ELA = 63 schools, 374 teachers, and 13,962 students for the treatment group; 63 schools, 394 teachers, and 15,423 students for the control group. Sample size for Year 2 mathematics = 63 schools, 389 teachers, and 14,186 students for the treatment group; 63 schools, 396 teachers, and 15,809 students for the control group.

The analyses were based on a three-level regression (students nested within teachers within schools) controlling for random assignment blocks and student background characteristics.

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

See appendix exhibits J.26 for additional details and J.28 and 29 for results by district.

SOURCE: District Administrative Records.

As we did for classroom practice and principal leadership, we explored whether the impact on achievement varied across districts, which might have occurred because of differences in district context or policy. We found that there was no statistically significant variation in the impact on achievement across the eight study districts in either reading/ELA or mathematics, in either Year 1 or 2. (See appendix exhibits J.28 and J.29.) In reading/ELA, the Year 2 impacts ranged from an effect size of -0.12 to 0.12 across the eight study districts, with none statistically significant. In mathematics, the Year 2 impacts ranged from -0.02 to 0.14, with none statistically significant.

As discussed above, we did find statistically significant variation across districts in the impact on classroom practice and principal leadership, but the results do not identify specific districts with consistently large or small effects for classroom practice, leadership, and achievement.

Thus, overall the results do not appear to suggest that the intervention worked particularly well in some districts, but poorly in others.

To parallel the separate analyses of impact on classroom practice and principal leadership in CLASS and FFT districts, we also examined the impact on achievement separately in the two sets of districts.¹²⁶ The magnitudes of the impacts were similar in the two sets of districts, although the impact of the intervention in mathematics in Year 1 was statistically significant only in the FFT districts.¹²⁷ (See exhibit 3.13.)

¹²⁶ Like the analysis of the impact of the intervention on leadership in CLASS and FFT districts, this analysis was prompted by the impact results for classroom practice in the two sets of districts. It was not part of the study's original analysis plan and thus should be viewed as exploratory.

¹²⁷ Because the impacts on achievement in CLASS and FFT districts are estimated with error, the apparent differences in results in the two sets of districts could be due to chance. A test of the differential impact of the intervention in CLASS and FFT districts was statistically significant in one of the four outcomes (reading/ELA in Year 2). (See appendix exhibit J.27.) Thus, there is little evidence of a systematic difference in impact in CLASS and FFT districts.

Exhibit 3.13. Average reading/ELA and mathematics achievement in CLASS districts and FFT districts, by treatment status and year

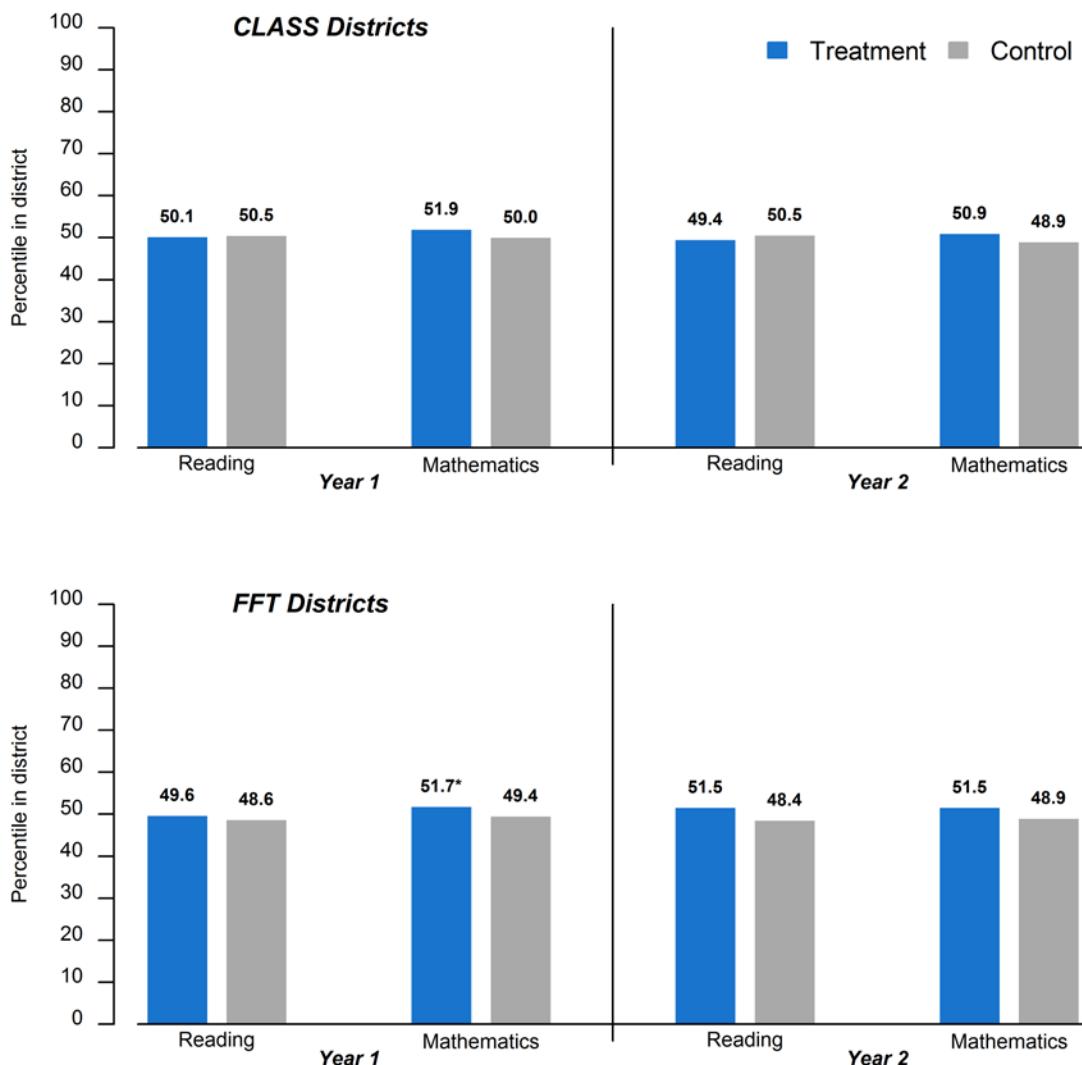


EXHIBIT READS: In CLASS districts, in Year 1, students in treatment schools received an average reading/ELA score at the 50.1st percentile in their district, compared to the 50.5th percentile for students in control schools..

NOTES: Year 1 sample size for reading/ELA in CLASS districts = 31 schools, 203 teachers, and 7,402 students for the treatment group; 32 schools, 240 teachers, and 8,447 students for the control group. Year 1 sample size for mathematics in CLASS districts = 31 schools, 232 teachers, and 8,269 students for the treatment group; 32 schools, 257 teachers, and 9,148 students for the control group. Year 1 sample size for reading/ELA in FFT districts = 32 schools, 181 teachers, and 5,732 students for the treatment group; 32 schools, 181 teachers, and 6,911 students for the control group. Year 1 sample size for mathematics in FFT districts = 32 schools, 179 teachers, and 5,698 students for the treatment group; 32 schools, 182 teachers, and 6,759 students for the control group. Year 2 sample size for reading/ELA in CLASS districts = 31 schools, 208 teachers, and 8,059 students for the treatment group; 32 schools, 231 teachers, and 8,997 students for the control group. Year 2 sample size for mathematics in CLASS districts = 31 schools, 230 teachers, and 8,315 students for the treatment group; 32 schools, 235 teachers, and 8,823 students for the control group. Year 2 sample size for reading/ELA in FFT districts = 32 schools, 166 teachers, and 5,903 students for the treatment group; 31 schools, 163 teachers, and 6,426 students for the control group. Year 2 sample size for mathematics in FFT districts = 32 schools, 159 teachers, and 5,871 students for the treatment group; 31 schools, 161 teachers, and 6,986 students for the control group. The analyses were based on a three-level regression (students nested within teachers within schools controlling for random assignment blocks and student background characteristics. * Difference between the treatment and control groups is statistically significant at the .05 level (two-tailed). See appendix exhibit J.27 for additional details.

SOURCE: District Administrative Records.

Association Among Classroom Practice, Leadership, and Achievement

The results above indicate that the performance feedback provided had a positive impact on some aspects of teacher classroom practice, principal leadership, and student achievement, consistent with the theory of action outlined in chapter 1. A remaining question is whether classroom practice, leadership, and achievement are linked as suggested by the theory: Did the impact of performance feedback on achievement occur by improving teachers' classroom practice? Or did the impact on achievement occur by improving principals' leadership, which could improve instructional focus, morale, and other factors related to achievement?

The study design does not permit a rigorous causal analysis addressing these questions. But the data we have permit us to explore whether teachers' classroom practice, using the study's outcome measure based on video-recorded lessons, was associated with their students' reading and mathematics achievement, controlling for students' prior achievement and other student and teacher background characteristics, which would be expected according to the theory of action. (See appendix H for a description of the methods, and appendix exhibits J.37a–c for more detail on the results for reading/ELA, and J.37d–f for mathematics.) We found an association of 0.06 between classroom practice as measured by the CLASS in Year 2 and students' mathematics achievement in Year 2—implying that students in classes taught by teachers with classroom practice that was a standard deviation above average would have achievement 0.06 standard deviations above average. This corresponds to about five weeks of learning.¹²⁸ Similarly, we found an association of 0.07 between classroom practice as measured by the FFT in Year 2 and mathematics achievement in Year 2. This corresponds to about seven weeks of learning.¹²⁹ These results are consistent with the theory of action, in that they suggest that performance feedback could have boosted achievement, in part, by improving teachers' classroom practice.

We conducted a similar analysis of the association between principal leadership and achievement in both Year 1 and Year 2, testing whether principals' leadership was associated with their students' reading and mathematics achievement, controlling for students' prior achievement and other student and principal background characteristics. We did not find a statistically significant association between either of the two measures principal leadership used in the study (teacher-principal trust or instructional leadership) and students' mathematics or reading achievement in either Year 1 or 2.¹³⁰ These results are not consistent with the theory of action, suggesting that improved leadership may not have been a factor in improved achievement.

¹²⁸ According to Hill et al. (2008), the average annual gain in mathematics is about 0.42 standard deviations for students in grades 4–8. A teacher with CLASS scores one standard deviation above average is predicted to have students 0.06 standard deviations above average, which translates into about $0.07/0.42 = 0.14$ of a year's achievement gain. Assuming a 36-week school year, this implies that the impact corresponds to five weeks of learning.

¹²⁹ Parallel analyses for reading indicate an association of 0.03 between classroom practice as measured by the CLASS in Year 2 and students' reading achievement, as well as an association of 0.03 for classroom practice as measured by the FFT.

¹³⁰ In other studies, these measures have been found to be associated with achievement. See, for example, Sebastian and Allensworth (2012).

Summary

The performance feedback tested in the study was intended to identify educators who needed support and to signal specific areas of practice for improvement. This chapter reported on treatment-control differences in the amount of performance feedback educators received; the impact of the intervention on educators' interest in improving their practice and their self-perceived effectiveness; and the impact on teachers' classroom practice, principals' leadership, and students' achievement. As intended, treatment teachers received more feedback on their classroom practice and more student growth information than control teachers, although they received less achievement information on individual students they taught. The oral feedback based on classroom observations received by treatment teachers was of longer duration and more likely to include ratings and written narrative information than the feedback received by control teachers. These findings suggest that treatment teachers received more frequent feedback with ratings than control teachers, as intended. Treatment principals also reported receiving more and longer instances of oral feedback that included ratings than control principals.

However, the intervention did not have most of the impacts on teachers' initial outcomes anticipated by the theory of action. Treatment teachers discussed topics covered on the CLASS and FFT instruments with the individuals providing feedback more frequently than control teachers. But despite this, in general they were no more likely than control teachers to indicate that they wanted to improve in these areas, and were no more likely to participate in professional development that covered these areas, than control teachers. The feedback treatment teachers received also did not lower their ratings of their own effectiveness in boosting student achievement in reading/ELA and mathematics. Principals were no more likely than control teachers to discuss topics covered on the VAL-ED. And like treatment teachers, they were no more likely to indicate they wanted to improve in areas covered by the VAL-ED, or to report that they participated in professional development that covered these areas. They also did not change their perceptions of their effectiveness as leaders in response to the feedback.

Although the intervention did not have the impacts on educators' initial outcomes anticipated by the theory of action, it did have an impact on aspects of the three main outcomes it was expected to affect. In particular, it had a positive impact on teachers' classroom practice as measured by the CLASS, but not as measured by the FFT. The impact on classroom practice occurred only in districts that used the CLASS; there was no effect in districts that used the FFT. The intervention also had a positive impact on both measures of principal leadership—instructional leadership and teacher-principal trust. Finally, in Year 1, the intervention had a positive impact on students' achievement in mathematics, amounting to about four weeks of learning. In Year 2, it had an impact on mathematics achievement that was similar in magnitude but not statistically significant. It did not have an impact on reading/ELA achievement in either year.

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

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., Gregory, A., Mikami, A.Y., Lun, J., Hamre, B.K., and Pianta, R.C. (2013). Observations of Effective Teaching in Secondary School Classrooms: Predicting Student Achievement With the CLASS-S. *School Psychology Review*, 42(1): 76–98.
- 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.
- 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://collegegated.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.
- Chetty, R., Friedman, J.N., and Rockoff, J.E. (2014a). Measuring the Impacts of Teachers II: Evaluating Bias in Teacher Value-Added Estimates. *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.
- Chicago Consortium of School Research (CCSR). (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>.
- 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.
- 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.
- 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>.
- Donaldson, M.L., and Papay, J.P. (2014). Teacher Evaluation Reform: Policy Lessons for School Principals. *Principal's Research Review*, 9(5): 1–8.
- 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.
- 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.
- Hill, C.J., Bloom, H.S., Black, A.R., and Lipsey, M.W. (2008). Empirical Benchmarks for Interpreting Effect Sizes in Research. *Child Development Perspectives*, 2 (3): 172–177.
- 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., 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.
- 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.
- Papay, J. (2012). Refocusing the Debate: Assessing the Purposes and Tools of Teacher Evaluation. *Harvard Educational Review*, 82(1): 123–141.
- 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.
- Rowan, B., and Raudenbush, S. (2016). Teacher Evaluation in American Schools. In D. Gitomer and C.A. Bell (Eds.), *Handbook of Research on Teaching* (5th ed.) (pp. 1159–1216). Washington, DC: American Educational Research Association.
- 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. (2015). Does Teacher Evaluation Improve School Performance? Experimental Evidence from Chicago's Excellence in Teaching Project. *Education Finance and Policy* 10(4): 1–38.
- 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 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.
- Wayne, A.J., Garet, M.S., Brown, S., Rickles, J., Song, M., and 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.
- 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.
- Whitehurst, G.J., Chingos, M.M., and Lindquist, K.M. (2014). *Evaluating Teachers With Classroom Observations: Lessons Learned in Four Districts*. Washington, DC: The Brookings Institution.

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

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. The third section presents the student, teacher, and principal turnover across the first and second year impact samples.

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 analyses were based on data for the baseline year (i.e., the year prior to the intervention). 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 of elementary schools in the study sample, elementary schools in similarly sized districts, and the national population, baseline year

School characteristic	Study sample	Elementary schools in 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

NOTES: "Similarly sized districts" are districts with at least 20 elementary and middle schools. Percentages for characteristics with multiple categories may not sum to 100 due to rounding. Differences between study schools and schools in similarly sized districts or the national population were tested using *t* tests. * Differences between study schools and schools in similarly sized districts or the national population is statistically significant at the .05 level (two-tailed).

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, baseline year

School characteristic	Study sample	Middle schools in 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

NOTES: "Similarly sized districts" are districts with at least 20 elementary and middle schools.

Percentages for characteristics with multiple categories may not sum to 100 due to rounding.

Differences between study schools and schools in similarly sized districts or the national population were tested using *t* tests.

* Differences between study schools and schools in similarly sized districts or the national population is statistically significant at the .05 level (two-tailed).

SOURCE: 2011–12 Common Core of Data.

Exhibit A.3. Background characteristics for schools in CLASS and FFT districts, baseline year

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

NOTES: Percentages for characteristics with multiple categories may not sum to 100 due to rounding.

† Reporting standards not met; in one or more cells, there are too few cases to report.

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. The analyses for schools and students were based on data for the baseline year; the analyses for principals and teachers were based on data for the fall of Year 1. The results using all eight study districts, as well as results for the CLASS and FFT districts separately, appear in exhibits A.4a–j.¹³¹

Exhibit A.4a. School background characteristics, by treatment status, baseline year

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 or 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		

NOTES: The analyses were based on a school-level regression model controlling for random assignment blocks.

None of the differences between the treatment and the control groups are statistically significant at the .05 level (two-tailed).

SOURCE: 2011–12 Common Core of Data.

Exhibit A.4b. School background characteristics in CLASS districts, by treatment status, baseline year

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 or 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		

NOTES: The analyses were based on a school-level regression controlling for random assignment blocks.

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

SOURCE: 2011–12 Common Core of Data.

¹³¹ Appendix exhibits J.1–13 provide baseline equivalence results for the Year 1 and 2 teacher and student impact samples.

Exhibit A.4c. School background characteristics in FFT districts, by treatment status, baseline year

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 or 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		

NOTES: The analyses were based on a school-level regression model controlling for random assignment blocks.

None of the differences between the treatment and the control groups are statistically significant at the .05 level (two-tailed).

SOURCE: 2011–12 Common Core of Data.

Exhibit A.4d. Principal background characteristics, by treatment status, fall of Year 1

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		

NOTES: The analyses were based on a principal-level regression model controlling for random assignment blocks.

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

† Reporting standards not met; in one or more cells, there are too few cases to report.

SOURCE: Fall 2012 District Administrative Records.

**Exhibit A.4e. Background characteristics of principals in CLASS districts,
by treatment status, fall of Year 1**

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		

NOTES: The analyses were based on a principal-level regression model controlling for random assignment blocks.

None of the differences between the treatment and the control groups are statistically significant at the .05 level (two-tailed).

† Reporting standards not met; in one or more cells, there are too few cases to report.

SOURCE: Fall 2012 District Administrative Records.

**Exhibit A.4f. Background characteristics of principals in FFT districts,
by treatment status, fall of Year 1**

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		

NOTES: The analyses were based on a principal-level regression model controlling for random assignment blocks.

None of the differences between the treatment and the control groups are statistically significant at the .05 level (two-tailed).

SOURCE: Fall 2012 District Administrative Records.

Exhibit A.4g. Teacher background characteristics, by treatment status, fall of Year 1

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 grade 4–8 teachers	575	594		

NOTES: 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 are statistically significant at the .05 level (two-tailed).

SOURCES: Fall 2012 District Administrative Records.

Exhibit A.4h. Background characteristics of teachers in CLASS districts, by treatment status, fall of Year 1

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 grade 4–8 teachers	337	344		

NOTES: 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 are statistically significant at the .05 level (two-tailed).

SOURCE: Fall 2012 District Administrative Records.

**Exhibit A.4i. Background characteristics of teachers in FFT districts,
by treatment status, fall of Year 1**

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 grade 4–8 teachers	238	250		

NOTES: The analyses were based on a two-level regression (teachers within schools) controlling for random assignment blocks.

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

SOURCE: Fall 2012 District Administrative Records.

Exhibit A.4j. Student background characteristics, by treatment status, baseline year

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 Grade 4–8 reading achievement	-0.029	0.022	-0.051	.111
2011–12 Grade 4–8 mathematics achievement	-0.009	-0.006	-0.003	.932
Number of grade 4–8 students	15,551	17,308		

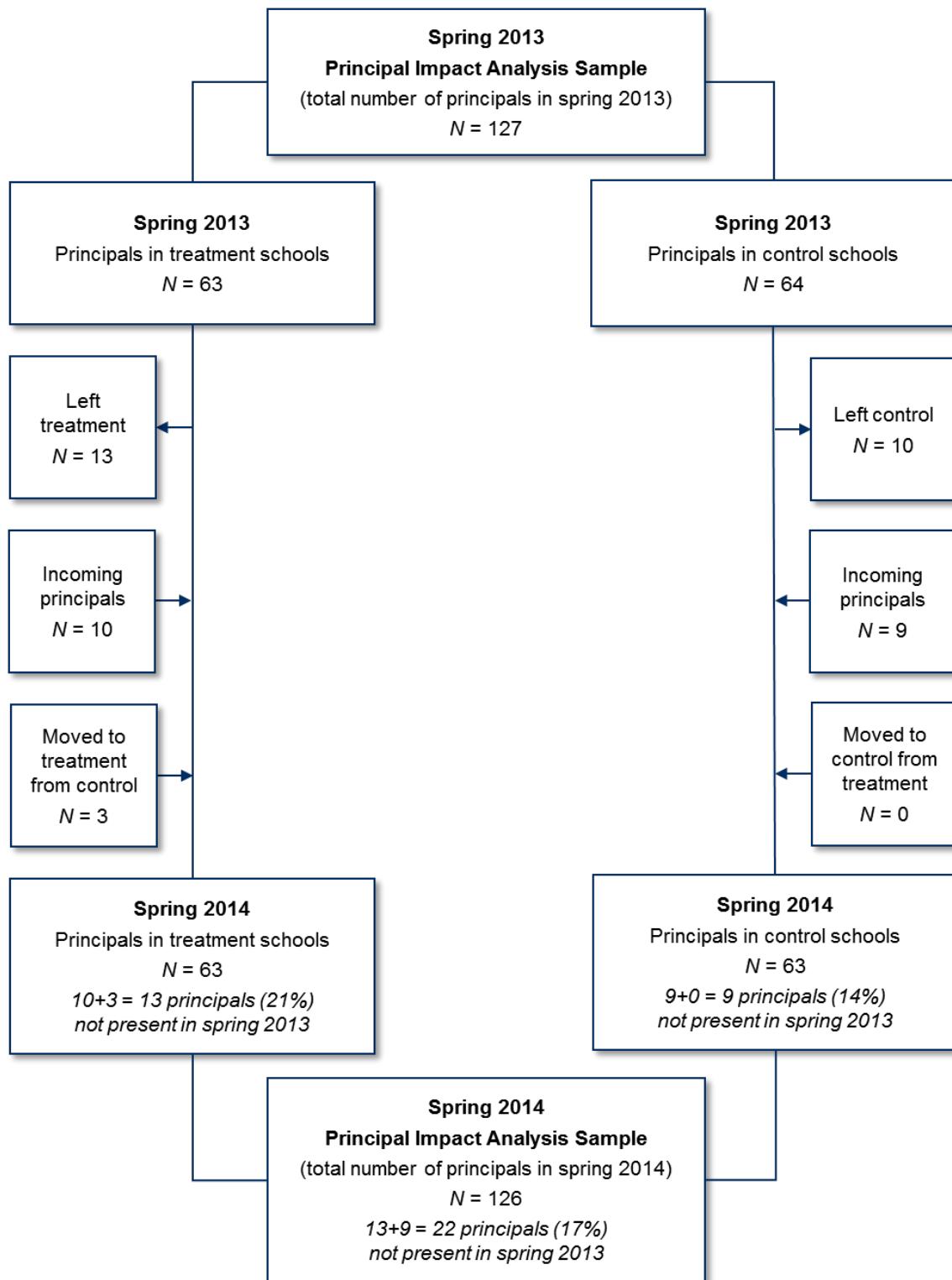
NOTES: The analyses were based on a three-level regression (students within teachers within schools) controlling for random assignment blocks.

None of the differences between the treatment and the control groups are statistically significant at the .05 level (two-tailed).

SOURCES: Spring 2012 District Administrative Records.

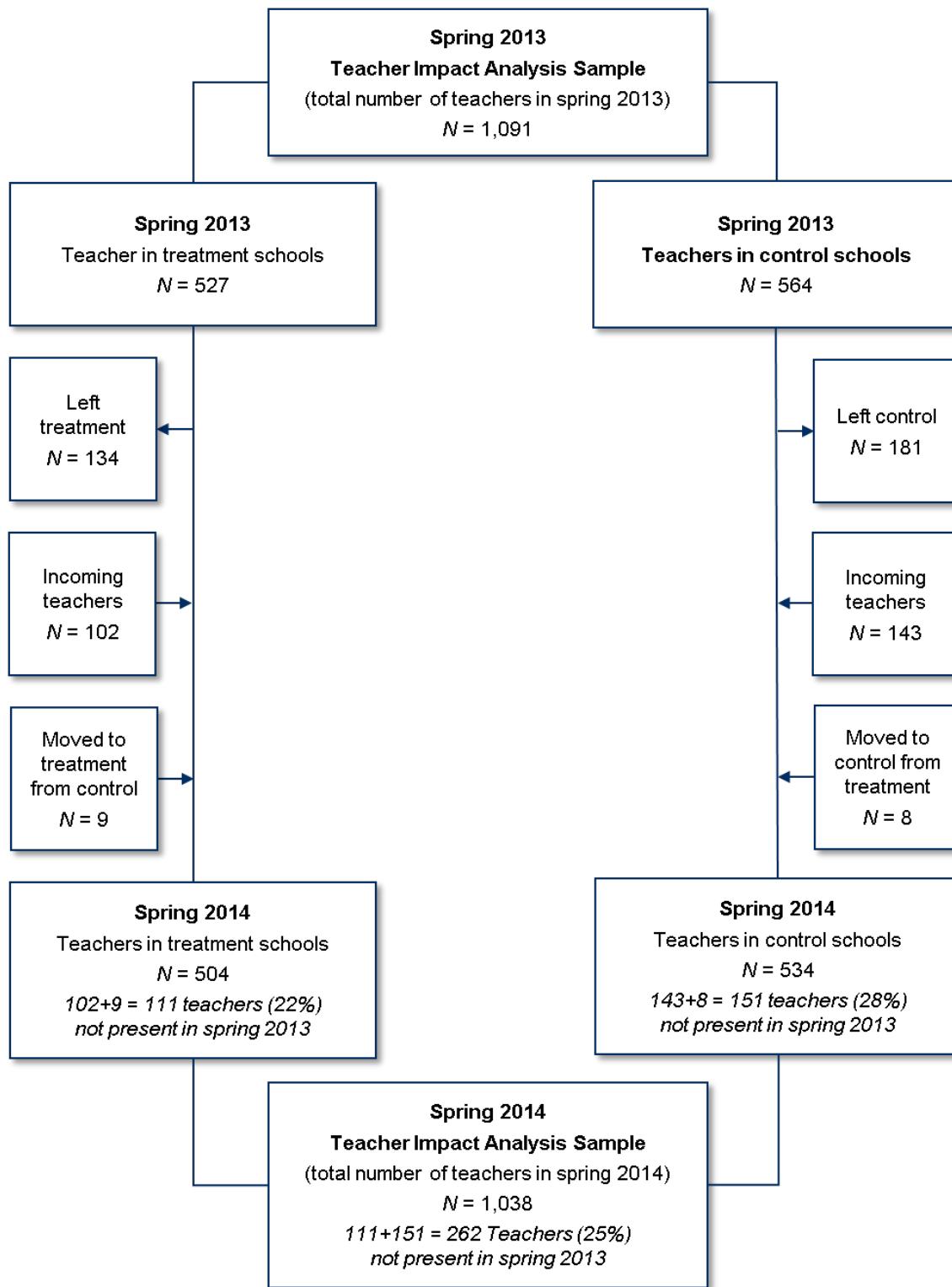
Sample Turnover Across Study Years

Exhibit A.5. Principal turnover across study years



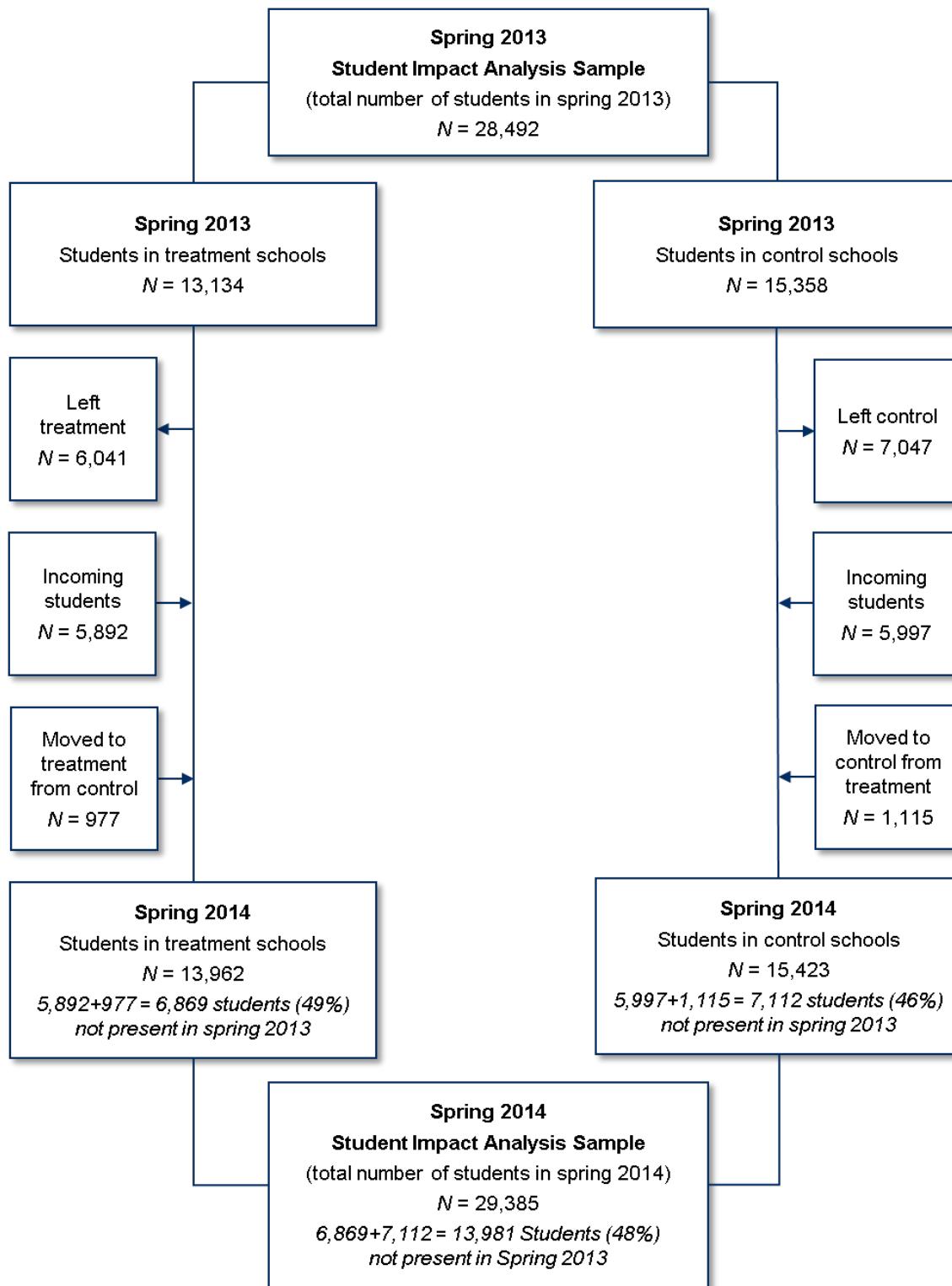
SOURCE: Study Records.

Exhibit A.6. Teacher turnover across study years



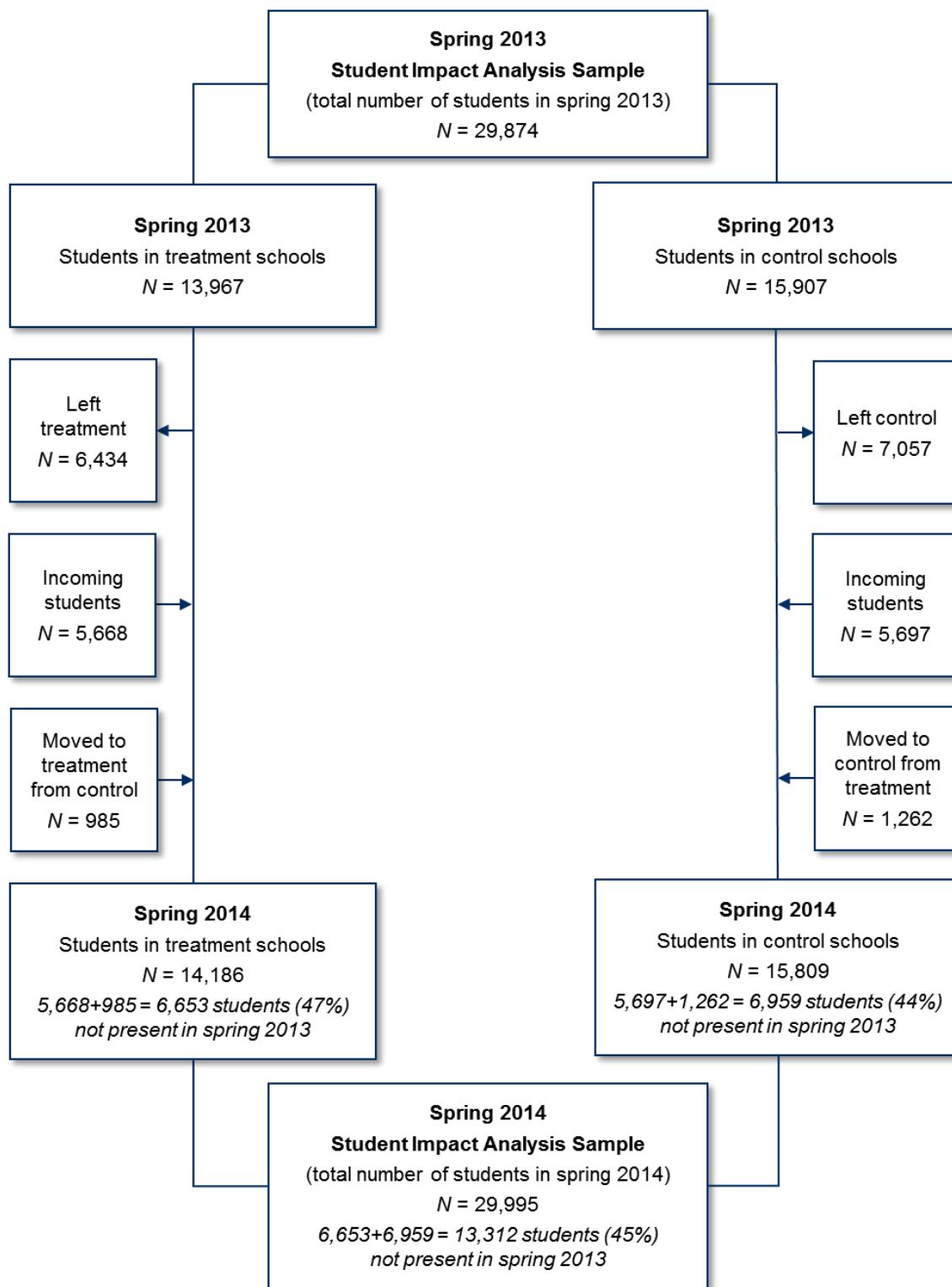
SOURCE: Study Records.

Exhibit A.7. Student turnover across study years, reading/ELA achievement impact sample



SOURCE: Study Records.

Exhibit A.8. Student turnover across study years, mathematics achievement impact sample



SOURCE: Study Records.

Exhibit A.9. Percentage of principals, teachers, and students who exited between spring Year 1 and 2, by treatment status

	Treatment group	Control group	Estimated difference	p value
Principals				
Exits (percentage)†	20.6	14.8	5.8	0.416
Teachers				
Exits (percentage)†	28.7	34.2	-5.5	0.053
Students				
Exits for reading/ELA achievement sample (percentage)†	23.6	21.5	2.1	0.314
Exits for mathematics sample (percentage)†	21.9	21.5	0.4	0.854

NOTES: Sample size for principals = 63 principals for the treatment group; 63 principals for the control group. Sample size for the grade 4–8 teachers in all districts = 63 schools and 527 teachers for the treatment group; 64 schools and 564 teachers for the control group. Sample size for students with reading scores = 63 schools, 239 teachers, and 8,016 students for the treatment group; 64 schools, 267 teachers, and 8,635 students for the control group.

The principal exit analysis was based on a principal-level regression controlling for random assignment blocks and principal background characteristics; the teacher exit analysis was based on a two-level analysis (teachers within schools) controlling for random assignment blocks and teacher background characteristics; student exit analyses were based on a three-level regression (students within teachers within schools) controlling for random assignment blocks and student background characteristics.

† Exiting principals, teachers, and students are defined consistently with exhibits A.5–8. Principals are defined as exiting if they left their school, unless they moved to another school in the study sample in the same condition. Similarly, teachers are defined as exiting if they left their school, unless they moved to another school in the study sample in the same condition. They are also defined as exiting if they moved to a grade or subject outside 4–8 reading/ELA and mathematics. Students are defined as exiting if they left their school, unless they moved to another school in the study sample in the same condition. Students in their schools' highest grade in Year 1 were excluded from the analysis, because they were required to leave their schools prior to Year 2.

The treatment group mean is the weighted average of the observed percentage exiting from treatment schools in each district, weighted by the number of treatment schools in the district. The control group mean is computed as the treatment group mean minus the estimated difference. Thus the treatment and control group means may not agree exactly with the values in appendix exhibits A.5–8.

None of the differences between the treatment and the control groups are statistically significant at the .05 level (two-tailed).

SOURCE: Study Records.

Realized Statistical Power for Impacts on Educator and Student Outcomes

We computed the minimum detectable effect size (MDES) based on the actual analysis sample and impact result for each primary outcome of the study. The realized MDESs range from 0.14 to 0.18 for classroom practice outcomes, from 0.26 to 0.29 for principal leadership outcomes, and from 0.05 to 0.09 for student achievement outcomes, as summarized in exhibit A.10.

Exhibit A.10. Realized minimum detectable effect sizes for educator and student outcomes, by year

Outcome	Year	Realized MDES
Classroom practice		
CLASS overall score	Year 2	0.18
FFT overall score	Year 2	0.14
Principal leadership		
Instructional leadership	Year 1	0.29
Teacher-principal trust	Year 1	0.26
Instructional leadership	Year 2	0.26
Teacher-principal trust	Year 2	0.28
Student achievement		
Reading	Year 1	0.05
Mathematics	Year 1	0.05
Reading	Year 2	0.06
Mathematics	Year 2	0.09

NOTE: Each realized MDES was computed based on the standard error of the impact estimate, standard deviation of the outcome in the control group, and the degrees of freedom for the impact analysis.

SOURCES: Spring 2014 Classroom Videos; Spring 2013 and Spring 2014 Teacher Surveys; District Administrative Records.

Appendix B. Details About Data Collection and Outcome Measures

This appendix provides details on the study's data collection activities and on its main outcome measures. The study team collected five types of data: data on the implementation of the intervention, including the intervention's ratings of educator performance; surveys of teachers and principals; videotapes of teacher classroom practice; and data on participant characteristics (which includes student achievement). After discussing the four types of data collected, we describe the main outcome measures: teacher classroom practice, principal leadership, and student achievement.

Data Collection

Data Collected on the Implementation of the Intervention

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

Exhibit B.1. Data collection schedule for intervention implementation data in each study year

Data	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 to record 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's online system.

Study-Hired Observer Questionnaire. At the end of the first and second years 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 grade 4–8 mathematics and reading/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 and 2014 with officials in each school district who were responsible for teacher and principal performance management. These interviews, each lasting approximately 90 minutes, sought contextual information regarding the districts' human resources policies (i.e., business as usual), primarily focusing on their teacher and principal evaluation system policies and the ways in which performance data were used. The interviews also collected information about the integration of the study's intervention with existing district processes.

Surveys of Teachers and Principals

In the spring of each study year, we administered surveys to teachers and principals in the study schools. The surveys focused on educators' experiences with performance evaluation and their initial outcomes. (These terms are discussed in chapter 1, in the section titled "Theory of action and research questions.") The teacher survey additionally included measures of principal leadership. The surveys for the treatment and control groups were identical, except that the treatment group surveys contained additional items asking about perceptions of the intervention. Specifically, the surveys for treatment principals in Year 1 and Year 2 asked about perceptions of the intervention's classroom practice measure. The surveys for treatment teachers and principals in Year 2 asked about perceptions of the intervention's student growth measure. These surveys also asked about perceptions of the classroom practice measure and principal leadership measure, respectively, compared to what was received from the district prior to the study.

The teacher survey was administered to all K–8 teachers of mathematics and reading/ELA; it took about 30 minutes to complete. For the teachers who were the focus of the study

(i.e., grade 4–8 teachers responsible for instruction in reading/ELA and mathematics), the response rate was 99.3 percent in the first year and 98.6 percent in the second year, as shown in exhibit B.2.

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 in the first year and 96.0 percent in the second year, as shown in exhibit B.2.

Exhibit B.2 Response rates for teacher survey, principal survey, and video-recording, overall and by treatment status

	Overall	Treatment group	Control group
Teacher survey ^a			
Year 1	99.3%	99.6%	98.9%
Year 2	98.6%	99.0%	98.1%
Principal survey			
Year 1	96.9%	96.8%	96.9%
Year 2	96.0%	96.8%	95.2%
Videotaping			
Year 2	91.6%	86.1%	96.8%

NOTE: ^aTeacher survey response rates are for the teachers who were the focus of the study (i.e., grade 4–8 teachers responsible for instruction in reading/ELA and mathematics).

Data Collected on Teacher Classroom Practice

To measure the impact of the intervention on classroom practice, we collected video recordings of treatment and control teachers in the spring of the second study year. These data were collected independent of the study's intervention. We video-recorded one lesson per teacher and then selected a random sample of half of the respondents for a second round of recording.¹³² Each recording captured approximately 30 consecutive minutes. The combined response rate for the video collection was 91.6 percent, with 86.1 percent for treatment teachers and 96.8 percent for control teachers, as shown in exhibit B.2.

The videographers were instructed to record a reading/ELA or mathematics lesson. For elementary teachers, we allowed recording of instruction in other topics if the videographer thought that waiting for instruction in reading/ELA or mathematics would disrupt the schedule for filming other teachers.

Data Collected on Participant Characteristics and Student Achievement

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

¹³² We video recorded two lessons for some teachers and one for others to achieve the desired precision while minimizing cost and burden.

characteristics of principals, teachers, and students in study schools from district administrative records in the summer and fall of 2012.

We collected additional district administrative records in fall 2013 and fall 2014, including individual student achievement records based on state tests in mathematics and reading/ELA that were administered in the spring of each year. Student achievement records were used to determine the impact of the intervention on student achievement at the end of the first and second study years. Student achievement records from spring 2012 were used as a covariate in analyses of the impact of the intervention on student achievement, as described in appendix H.

Main Outcome Measures

This section discusses the study's main outcome measures: teacher classroom practice, principal leadership, and student achievement.

Outcome Measures for Teacher Classroom Practice

The outcome measure for teacher classroom practice was based on videotapes that were recorded independently from the intervention. All videos were coded using CLASS and FFT, forming the study's two outcome measures for teacher classroom practice.

The study team divided each 30-minute video into two 15-minute segments, and randomly selected either the first or second segment for coding. Focusing on one 15-minute segment was intended to balance the costs of coding videos with the need for precise measures of classroom practice. In a study using FFT to code videos, Ho and Kane (2013) found that focusing coders on the first 15 minutes of instruction produced mean FFT scores similar to those obtained from coding the full 30 minutes of instruction, but with some loss of precision.

To remove rater effects from impact analyses, coders were assigned equal numbers of treatment and control videos. When feasible, these videos were drawn from the same random assignment block. Finally, to avoid influencing the study results, the videos and scores were kept confidential from the study participants and the study's implementation team.

There were two separate groups of coders: one for CLASS and one for FFT. All coders received the standard training for their instrument and passed the observer certification test. During the coding work, the coders were required to participate in calibration exercises approximately once every three weeks. In the exercises, coders watched videos and coded them, much like the observer certification tests, and could attend follow-up discussions about the correct scores for each video. In addition, each coder's workload included some videos that had already been master-coded by Teachstone and Danielson Group. These were used to monitor coders' performance; coders were not told which videos were being used to test the accuracy of their ratings. Repeatedly failing calibration exercises or incorrectly coding the master-coded segments was a basis for follow-up training and in some cases discontinuing the use of a coder.

Each measure was formed by computing the mean of the responses to the items, as is conventionally done.

Outcome Measures for Principal Leadership

To provide a common measure of principal leadership in treatment and control schools, we relied on teachers' responses to survey items designed to assess principal leadership. The items appeared on the teacher survey we administered to treatment and control teachers in the spring of each year.

The survey items were adapted from a set of items on the teacher survey of the Chicago Consortium on School Research (CCSR 2012), which were shown to have an association with the quality of instruction and student achievement (Sebastian and Allensworth 2012).¹³³ Sebastian and Allensworth (2012) explain that the leadership items capture two constructs discussed in the literature: *instructional leadership* and *teacher-principal trust*. The first is intended to capture teachers' perceptions of the principal's leadership related to teaching and learning (e.g., to what extent the principal sets high standards for teaching and learning). The items on teacher-principal trust scale are intended to capture the teacher's perception that the principal is trustworthy (e.g., that the principal places the needs of children ahead of personal interests).

We used eight items to measure instructional leadership and five to measure teacher-principal trust. (See exhibit B.3.) In response to each item, respondents could choose on a 1 to 4 scale. Each measure was formed by computing the mean of the responses to the items. The reliabilities (Cronbach's alpha) of the principal instructional leadership scale and teacher-principal trust scales were 0.95 and 0.92.

¹³³ It was not feasible to use the VAL-ED itself as an outcome measure. By the time of the Year 2 spring surveys, a large majority of treatment teachers had already completed the VAL-ED four times, making it likely that they would respond to the survey with a disposition or framework different from that used by control teachers, who had never before completed a VAL-ED survey.

Exhibit B.3. Item composition and reliabilities of principal leadership scales

Scale	Items
<u>Principal instructional leadership</u> Scale: disagree strongly, disagree somewhat, agree somewhat, agree strongly Year 1 Cronbach's alpha = 0.96 Year 2 Cronbach's alpha = 0.95	Makes clear to the staff his or her instructional expectations. Communicates a clear vision for our school. Sets high standards for teaching. Understands how children learn. Sets high standards for student learning. Encourages teachers to implement what they have learned from their professional development. Actively tracks student academic progress. Actively monitors the quality of teaching in this school.
<u>Teacher-principal trust</u> Scale: disagree strongly, disagree somewhat, agree somewhat, agree strongly Year 1 Cronbach's alpha = 0.93 Year 2 Cronbach's alpha = 0.92	It's OK in this school to discuss worries and frustrations with the principal. The principal takes a personal interest in the professional development of teachers. The principal is aware of areas in which I would like to improve. The principal is responsive to teachers' input. The principal places the needs of children ahead of personal interests.

Outcome Measures for Student Achievement

The study took place in five states, each of which used different assessments for state accountability testing. To form common metrics of student achievement in reading/ELA and mathematics across the study districts, we standardized students' scores separately in each state, based on the state mean and standard deviation for each of the two subjects.

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.

The reliability estimates presented in this appendix are based on the performance information generated by the intervention in Year 2. A parallel appendix in the study’s first report has estimates based on Year 1. A summary of the reliabilities for Year 1 and Year 2 is provided in exhibit C.1.

Exhibit C.1. Summary of reliability estimates for measures of educator performance

Measure	Year 1	Year 2
Classroom observation overall scores		
CLASS single-window score	.24	.33
FFT single-window score	.49	.51
CLASS four-window average score ^a	.42 to .50	.53 to .61
FFT four-window average score ^a	.69 to .75	.70 to .77
Classroom observation dimension score differences		
CLASS single-window score	.19	.19
FFT single-window score	.09	.12
CLASS four-window average score ^a	.35 to .43	.35 to .43
FFT four-window average score ^a	.18 to .23	.24 to .30
Value-added		
Reading score	.44	.46
Mathematics score	.68	.67
Subject-score differences	.52	.50
VAL-ED overall score		
Fall	.19	.32
Spring	.51	.49
VAL-ED dimension score differences		
Fall core components	.36	.48
Fall key processes	.29	.31
Spring core components	.50	.43
Spring key processes	.20	.07

NOTE: ^a The range of reliabilities for classroom observations are based on assumptions about the proportion of within-teacher variance (error variance) due to observers rather than occasions, with the reported range based on 25–75 percent of the between-teacher variance due to observers.

SOURCES: Teachstone Online System; Teachscape Online System; AIR Value-Added System; VAL-ED Surveys.

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

Although these dimensions of variation are typically viewed as sources of error in analyses of reliability, the second—systematic differences across occasions—may reflect at least some true variation in the context of the study of teacher and leader feedback. In particular, according to the theory of action underlying the intervention, teachers may improve their practice from one observed lesson to the next. In the following section, we briefly discuss this issue as it pertains to the study’s measures of classroom practice.

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, as described above, 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.2 for CLASS and exhibit C.3 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. 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.

Because 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. In the right-side columns of exhibits C.2 and C.3, 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 results shown in exhibits C.2 and C.3 are based on the assumption that all seven sources of variation listed above are error. But as discussed earlier, at least some of the second source (systematic occasion variance) might reflect true improvement. We do not have a definitive way to assess the magnitude of the variation in ratings across occasions due to true improvement. But one plausible source of information is the observed trend in means from one occasion to the next. Exhibit D.7c provides the CLASS and FFT mean scores for each of the four observation waves each year. For example, the four CLASS means for Year 2 are 5.30, 5.46, 5.63, and 5.81, and the variance among the four means is 0.036. This variation in average ratings across the four windows could be due at least in part to true teacher improvement in practice. According to exhibit C.2, the within-teacher variance (error variance) for CLASS in Year 2 is 0.35. This is the variation due to observers and occasions. The variation potentially due to improvement is about

10 percent of the within-teacher variance (0.35). Taking it into account would result in only a modest increase in the estimated reliability.¹³⁴

The Year 2 reliability estimates presented in exhibits C.2 and C.3 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 .33 and .51 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).¹³⁵ 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.

¹³⁴ The occasion variance for FFT Year 2 is 0.003, or about 4 percent of the within-teacher variance (error variance) of 0.07.

¹³⁵ 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 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.2. Estimated reliabilities for CLASS overall scores and dimension scores, Year 2

CLASS 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	0.17	0.35	.33	.67	.66	.61	.57	.53	.50
Domain: Emotional support									
Positive climate	0.23	0.58	.28	.72	.61	.56	.51	.47	.44
Teacher sensitivity	0.27	0.61	.31	.69	.64	.58	.54	.50	.47
Regard for student perspectives	0.32	0.95	.25	.75	.58	.52	.48	.44	.41
Domain: Classroom organization									
Behavior management	0.13	0.45	.23	.77	.54	.49	.44	.40	.37
Productivity	0.13	0.45	.22	.78	.54	.48	.43	.40	.37
Negative climate (reverse coded)	0.01	0.08	.08	.92	.26	.22	.19	.17	.15
Domain: Instructional support									
Instructional learning formats	0.28	0.74	.28	.72	.60	.55	.50	.47	.43
Content understanding	0.34	0.88	.28	.72	.61	.55	.51	.47	.43
Analysis and inquiry	0.29	1.69	.15	.85	.41	.36	.32	.28	.26
Quality of feedback	0.31	1.07	.22	.78	.54	.48	.44	.40	.37
Instructional dialogue	0.39	1.18	.25	.75	.57	.51	.47	.43	.40
Domain: Student engagement	0.24	0.53	.31	.69	.65	.59	.55	.51	.48

NOTES: Sample size = 303 teachers.

^aThe proportion of between-teacher variance is also the reliability for ratings based on a single occasion and a single observer.

SOURCE: Teachstone Online System.

Exhibit C.3. Estimated reliabilities for FFT overall scores and dimension scores, Year 2

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	.51	.49	.81	.77	.74	.70	.68
Domain 2: Classroom environment									
Creating an environment of respect and rapport	0.11	0.22	.35	.65	.68	.63	.59	.55	.51
Establishing a culture for learning	0.09	0.19	.32	.68	.65	.60	.56	.52	.49
Managing classroom procedures	0.08	0.19	.30	.70	.64	.58	.54	.50	.47
Managing student behavior	0.10	0.22	.31	.69	.65	.59	.55	.51	.48
Domain 3: Instruction									
Communicating with students	0.12	0.20	.37	.63	.70	.65	.61	.57	.54
Using questioning and discussion techniques	0.09	0.21	.30	.70	.64	.58	.54	.50	.47
Engaging students in learning	0.11	0.21	.35	.65	.68	.63	.59	.55	.52
Using assessment in instruction	0.07	0.25	.23	.77	.54	.49	.44	.40	.37

NOTES: Sample size = 199 teachers.

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.

^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 Year 2 variance decomposition results and the reliability estimates for differences between dimension scores are presented in exhibit C.4 for CLASS and exhibit C.5 for FFT.

Exhibit C.4. Estimated variance components and reliabilities for dimension score differences for CLASS and FFT, Year 2

Source of variance	CLASS		FFT	
	Estimated variance component	Proportion of total variance	Estimated variance component	Proportion of total variance
teacher (<i>t</i>)	0.19	.11	0.07	.24
window (<i>w</i>)	0.05	.03	0.00	.01
dimension (<i>d</i>)	0.62	.37	0.01	.03
<i>t</i> × <i>w</i>	0.26	.16	0.05	.16
<i>t</i> × <i>d</i>	0.11	.07	0.02	.07
<i>w</i> × <i>d</i>	0.01	.01	0.00	.00
residual	0.44	.26	0.14	.49
Reliability estimates		CLASS	FFT	
Single-observation reliability		.19		.12
Four-window average reliability estimate				
0% observer error		.49		.35
25% observer error		.43		.30
50% observer error		.39		.27
75% observer error		.35		.24
100% observer error		.32		.21

NOTES: Sample size = 14,323 CLASS scores (303 teachers × 4 windows × 12 dimensions) and 7,452 FFT scores (199 teachers × 4 windows × 10 components). Not all teachers had scores for all windows and all dimensions/components.

SOURCES: Teachstone Online System and Teachscape Online System.

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.¹³⁶

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 Year 2 variance decomposition results for each subject are presented in exhibit C.5. 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:

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

¹³⁶ 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.5. Estimated reliabilities for value-added scores based on two years of student growth data, Year 2

Subject	Variance estimate		Proportion of variance		Reliability based on two years
	Between teacher	Within teacher	Between teacher ^a	Within teacher	
Reading	0.004	0.010	.30	.70	.46
Mathematics	0.021	0.021	.51	.49	.67

NOTES: Sample size = 974 teachers for reading; 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/ELA 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 Year 2 variance decomposition results and the reliability estimates for differences between subject value-added scores are presented in exhibit C.6. 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.6. Estimated variance components and reliability for subject-specific value-added score differences, Year 2

Source of variance	Estimated variance component	Proportion of total variance
Teacher (<i>t</i>)	0.01	.23
Year (<i>y</i>)	0.00	.00
Subject (<i>s</i>)	0.00	.00
<i>t</i> × <i>y</i>	0.01	.19
<i>t</i> × <i>s</i>	0.01	.19
<i>y</i> × <i>s</i>	0.00	.00
Residual	0.01	.39
Reliability estimate		.50

NOTES: Sample size = 2,696 value-added scores (674 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). Because 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 Year 2 variance decomposition results for the overall score and the dimension scores are presented in exhibit C.7 for fall and exhibit C.8 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.7. Estimated reliabilities for VAL-ED overall scores and dimension scores, fall of Year 2

VAL-ED dimension	Variance estimate		Proportion of variance		Reliability estimate
	Between principal	Within principal	Between principal	Within principal	
Overall score	0.04	0.23	.13	.87	.32
Core components					
High standards for student learning	0.05	0.23	.18	.82	.40
Quality instruction	0.05	0.24	.16	.84	.37
Culture of learning and professional behavior	0.04	0.24	.14	.86	.33
Connections to external communities	0.00	0.29	.01	.99	.03
Performance accountability	0.05	0.29	.15	.85	.35
Rigorous curriculum	0.04	0.24	.15	.85	.34
Key processes					
Planning	0.04	0.21	.15	.85	.35
Implementing	0.04	0.21	.17	.83	.38
Supporting	0.05	0.23	.19	.81	.41
Advocating	0.01	0.27	.05	.95	.14
Communicating	0.03	0.28	.11	.89	.27
Monitoring	0.03	0.27	.11	.89	.28

NOTE: Sample size = 63 principals.

SOURCE: Fall 2013 VAL-ED Surveys.

Exhibit C.8. Estimated reliabilities for VAL-ED overall scores and dimension scores, spring of Year 2

VAL-ED dimension	Variance estimate		Proportion of variance		Reliability estimate
	Between principal	Within principal	Between principal	Within principal	
Overall score	0.05	0.17	.24	.76	.49
Core components					
High standards for student learning	0.07	0.17	.28	.72	.54
Quality instruction	0.05	0.20	.20	.80	.42
Culture of learning and professional behavior	0.07	0.20	.26	.74	.51
Connections to external communities	0.03	0.20	.12	.88	.29
Performance accountability	0.07	0.20	.27	.73	.52
Rigorous curriculum	0.05	0.19	.22	.78	.46
Key processes					
Planning	0.06	0.17	.27	.73	.53
Implementing	0.05	0.19	.21	.79	.44
Supporting	0.06	0.18	.23	.77	.48
Advocating	0.05	0.19	.21	.79	.44
Communicating	0.06	0.19	.24	.76	.49
Monitoring	0.05	0.20	.19	.81	.41

NOTE: Sample size = 63 principals.

SOURCE: Spring 2014 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 inter-related ways: based on six core components and based on six key processes. Because 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 Year 2 variance decomposition results and the reliability estimates for differences between scores are presented in exhibit C.9 for the fall wave and exhibit C.10 for the spring wave. We conducted separate analyses for the core components and key processes.

Exhibit C.9. Estimated variance components and reliabilities for VAL-ED dimension score differences, fall of Year 2

Source of variance	Core components		Key processes	
	Estimated variance component	Proportion of total variance	Estimated variance component	Proportion of total variance
Principal (<i>p</i>)	0.03	.10	0.03	.12
Respondent group (<i>r</i>)	0.00	.00	0.00	.00
Dimension (<i>d</i>)	0.01	.03	0.00	.01
<i>p</i> × <i>r</i>	0.23	.75	0.23	.81
<i>p</i> × <i>d</i>	0.01	.03	0.00	.01
<i>r</i> × <i>d</i>	0.00	.01	0.00	.00
Residual	0.03	.08	0.02	.07
Reliability estimate	.48		.31	

NOTES: Sample size = 1,134 core component scores and 1,134 key process scores (63 principals × 3 respondent groups × 6 dimensions). Not all principals had scores from all respondent groups and all dimensions.

SOURCE: Fall 2013 VAL-ED Surveys.

Exhibit C.10. Estimated variance components and reliabilities for VAL-ED dimension score differences, spring of Year 2

Source of variance	Core components		Key processes	
	Estimated variance component	Proportion of total variance	Estimated variance component	Proportion of total variance
Principal (p)	0.06	.24	0.07	.26
Respondent group (r)	0.01	.04	0.01	.04
Dimension (d)	0.01	.04	0.00	.01
$p \times r$	0.16	.57	0.16	.61
$p \times d$	0.01	.03	0.00	.00
$r \times d$	0.00	.01	0.00	.00
Residual	0.03	.09	0.02	.08
Reliability estimate		.43		.07

NOTES: Sample size = 1,134 core component scores and 1,134 key process scores (63 principals \times 3 respondent groups \times 6 dimensions). Not all principals had scores from all respondent groups and all dimensions.

SOURCE: Spring 2014 VAL-ED Surveys.

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

Exhibit D.1. Percentage of treatment principals who agreed somewhat or strongly with each statement about the observations they conducted, by year

Statement	All districts	CLASS districts	FFT districts
Year 1			
I knew what to look for when I observed teachers.	100.0	100.0	100.0
I had a clear sense of what written feedback to give the teacher.	92.9	≥ 89.0 [†]	≥ 90.0 [†]
I had a clear sense of what oral feedback to give the teacher.	92.8	≥ 89.0 [†]	≥ 90.0 [†]
I had a clear sense of which score to give teachers on CLASS/FFT dimensions.	≥ 94.0 [†]	≥ 89.0 [†]	≥ 90.0 [†]
Number of principals (Year 1)	59	29	30
Year 2			
I had a clear sense of what kinds of things I was looking for when I observed teachers.	100.0	100.0	100.0
I had a clear understanding of how teachers should be scored on the CLASS dimensions/FFT components.	≥ 94.0 [†]	≥ 87.0 [†]	≥ 90.0 [†]
I had a clear sense of what written and verbal feedback to give the teacher.	≥ 94.0 [†]	≥ 87.0 [†]	≥ 90.0 [†]
On average, teachers performed better than I expected during the observations.	73.0	75.8	70.2
I have a clear idea of what the CLASS/FFT rating system views as "good instruction."	100.0	100.0	100.0
Number of principals (Year 2)	54	24	30

NOTE: [†] Reporting standards not met, too few cases report the exact percentage.

SOURCES: Spring 2013 and Spring 2014 Principal Surveys.

Exhibit D.2. Mean number of feedback sessions K–3 treatment teachers received, by year and in total

	All districts	CLASS districts	FFT districts
Year 1 feedback sessions (Year 1 impact sample)	1.8	1.7	2.0
Year 2 feedback sessions (Year 2 impact sample)	2.0	2.0	2.0
Year 1 and 2 feedback sessions (Year 2 impact sample)	3.4	3.3	3.5

NOTES: Sample size for Year 1 = 642 teachers (367 CLASS and 275 FFT). Sample size for Year 2 = 673 teachers (397 CLASS and 276 FFT).

SOURCES: Teachstone Online System and Teachscape Online System.

Exhibit D.3. Percentage of study-hired observers who reported that they engaged in a given activity in two-thirds or more of the feedback sessions they conducted, by year

	All districts	CLASS districts	FFT districts
Year 1			
I gave teachers a hard copy of their classroom observation report.	55.9	71.4	41.9
I showed teachers their classroom observation report on a computer/tablet.	46.6	29.6	61.3
Showed a video about dimensions that needed improvement.	32.2	†	†
Number of study-hired observers (Year 1)	58-59	27-28	31
Year 2			
I gave teachers a hard copy of their classroom observation report.	71.6	87.0	54.8
I showed teachers their classroom observation report on a computer/tablet.	51.2	43.2	59.5
Showed a video about dimensions that needed improvement.	34.1	56.5	9.5
Number of study-hired observers (Year 2)	86-88	44-46	42

NOTE: † Reporting standards not met; in one or more cells, there are too few cases to report.

SOURCES: Spring 2013 and Spring 2014 Study-Hired Observer Surveys.

Exhibit D.4. Percentage of study-hired observers who reported that teachers were engaged in the discussion in two-thirds or more of the feedback sessions they conducted, by year

	All districts	CLASS districts	FFT districts
Year 1			
Teachers were engaged in the discussion.	79.7	78.6	80.6
Number of study-hired observers (Year 1)	59	28	31
Year 2			
Teachers were engaged in the discussion.	92.0	93.3	90.5
Number of study-hired observers (Year 2)	87	45	42

SOURCES: 2013 and 2014 Study-Hired Observer Survey.

Exhibit D.5. Average percentage of teachers that study-hired observers felt needed significant or some help according to the CLASS or FFT instrument, Year 2

	All districts	CLASS districts	FFT districts
Percentage of teachers that study-hired observers felt needed <u>significant help</u> according to the CLASS or FFT instrument.	17.2	21.5	11.3
Percentage of teachers that study-hired observers felt needed <u>some help</u> according to the CLASS or FFT instrument.	45.8	50.8	39.1
Number of study-hired observers	76	44	32

SOURCE: Spring 2014 Study-Hired Observer Survey.

Exhibit D.6a. Distribution of K–3 teachers across performance levels based on CLASS overall scores in each observation window, and the two-window average in each year

	Ineffective	Developing effectiveness	Effective	Highly effective
Year 1				
Window 1	0.0	0.8	17.0	82.2
Window 2	0.0	†	†	87.5
Two-window average	0.0	†	†	88.8
Year 2				
Window 1	0.0	1.3	2.2	76.7
Window 2	0.0	†	†	81.8
Two-window average	0.0	0.0	18.5	81.5

NOTES: Sample size for Year 1 = 376 teachers in Window 1, 360 teachers in Window 2, 376 teachers for the two-window average. Sample size for Year 2 = 399 teachers in Window 1, 390 teachers in Window 2, 399 teachers for the two-window average. Percentages for each window and for the two-window average may not sum to 100 percent due to rounding.

† Reporting standards not met; in one or more cells, there are too few cases to report.

SOURCE: Teachstone Online System.

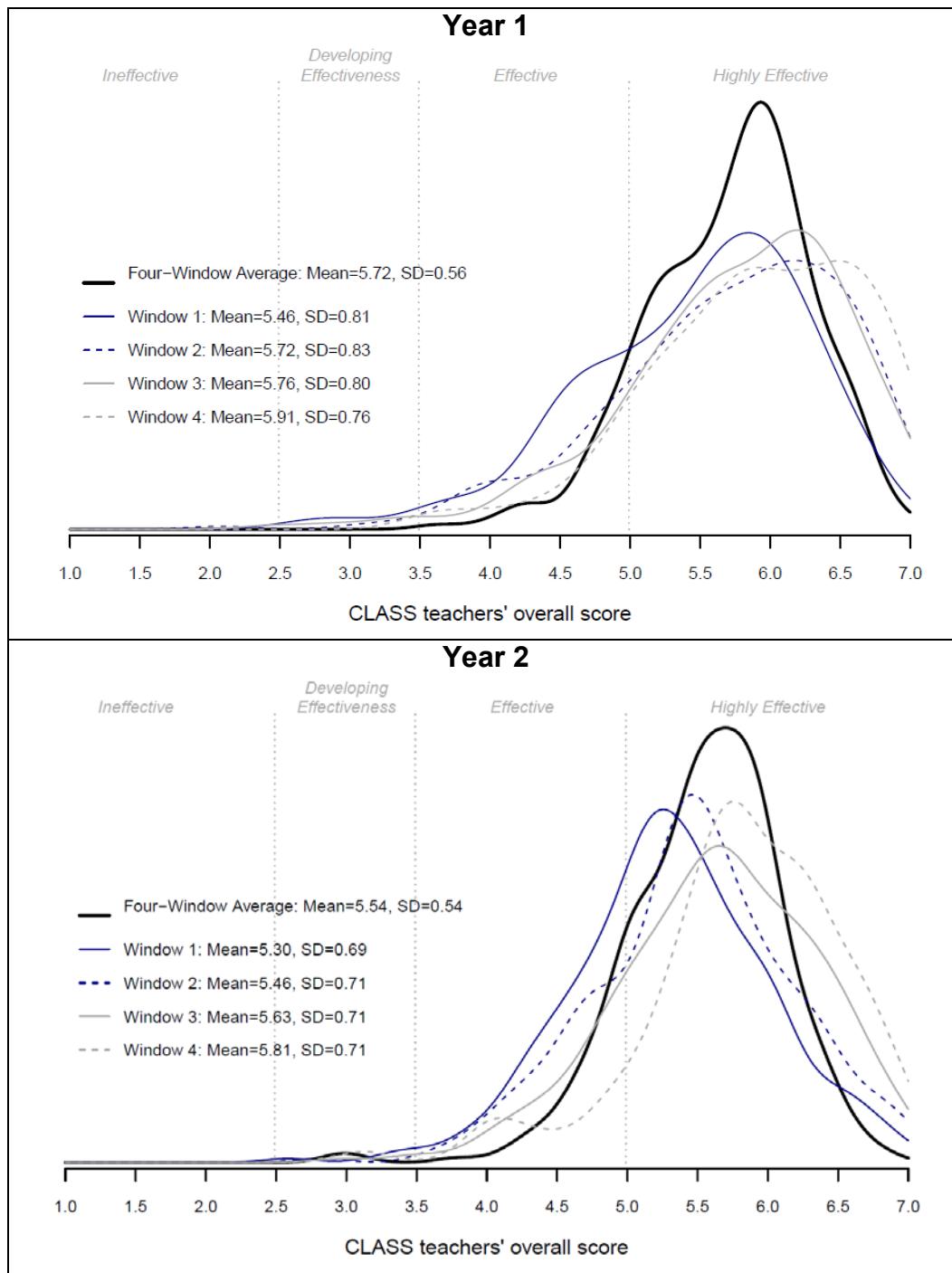
Exhibit D.6b. Distribution of K–3 teachers across performance levels based on FFT overall scores in each observation window, and the two-window average in each year

	Score of 1.00 to 1.49	Score of 1.50 to 2.49	Score of 2.50 to 3.49	Score of 3.50 to 4.00
Year 1				
Window 1	0.0	4.7	90.9	4.3
Window 2	0.0	5.1	84.1	8.0
Two-window average	0.0	3.3	94.2	2.5
Year 2				
Window 1	0.0	4.7	86.1	9.1
Window 2	0.0	2.2	86.3	11.5
Two-window average	0.0	2.2	92.3	5.5

NOTES: Sample size for Year 1 = 276 teachers in Window 1, 268 teachers in Window 2, 276 teachers for the two-window average. Sample size for Year 2 = 274 teachers in Window 1, 270 teachers in Window 2, 274 teachers for the two-window average. Percentages for each window and for the two-window average may not sum to 100 percent due to rounding.

SOURCE: Teachscape Online System.

Exhibit D.7a. Distribution of teachers based on their CLASS overall scores in each observation window and the four-window average, by year



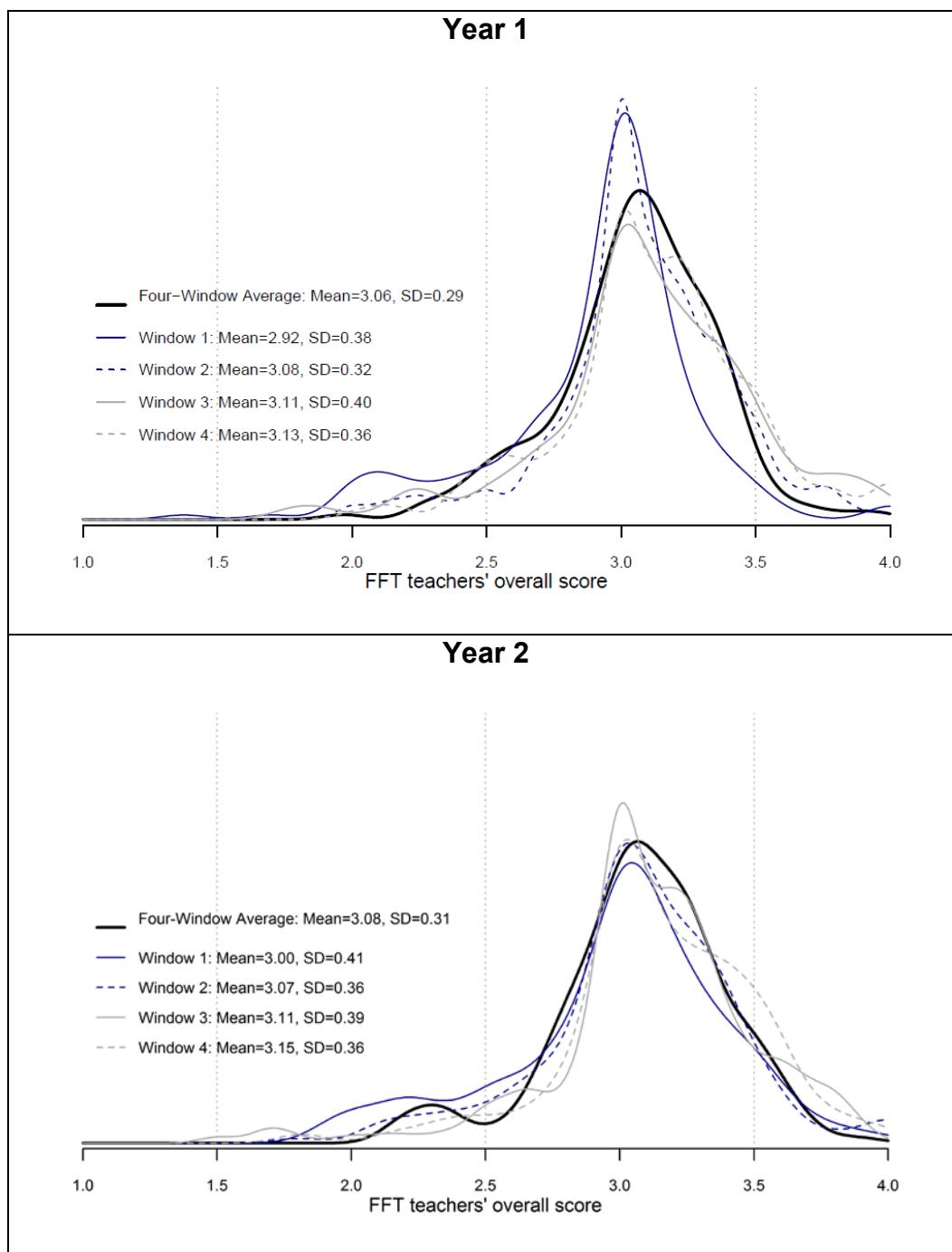
NOTES: Sample size for Year 1 = 265 observations in Window 1, 307 in Window 2, 309 in Window 3, and 283 in Window 4; Sample size for Year 2 = 297 observations in Window 1; 295 in Window 2; 300 in Window 3; and 302 in Window 4.

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.

See appendix exhibit D.5a for detailed information about the distribution of four-window average CLASS observation scores for K–3 teachers.

SOURCE: Teachstone Online System.

Exhibit D.7b. Distribution of teachers based on their FFT overall scores in each observation window and the four-window average, by year



NOTES: Sample size for Year 1 = 216 teachers in Window 1, 220 teachers in Window 2, 221 teachers in Window 3, 219 teachers in Window 4, and 222 teachers for the 4-Window average. Sample size for Year 2 = 191 teachers in Window 1, 196 teachers in Window 2, 196 teachers in Window 3, 198 teachers in Window 4, and 199 for the 4-window average. 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 gray 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. See appendix exhibit D.5b for detailed information about the distribution of four-window average FFT observation scores for K–3 teachers.

SOURCE: Teachscape Online System.

Exhibit D.7c. Descriptive statistics for CLASS and FFT overall scores in each observation window, by year

	<i>N</i>	Mean	Standard deviation
CLASS Year 1			
Window 1	265	5.46	0.81
Window 2	307	5.72	0.83
Window 3	309	5.76	0.80
Window 4	283	5.91†	0.76
CLASS Year 2			
Window 1	297	5.30*‡	0.69
Window 2	295	5.46*	0.71
Window 3	300	5.63	0.71
Window 4	302	5.81†	0.71
FFT Year 1			
Window 1	216	2.92	0.38
Window 2	220	3.08	0.32
Window 3	221	3.11	0.40
Window 4	219	3.13†	0.36
FFT Year 2			
Window 1	191	3.00*‡	0.41
Window 2	196	3.07	0.36
Window 3	196	3.11	0.39
Window 4	198	3.15†	0.36

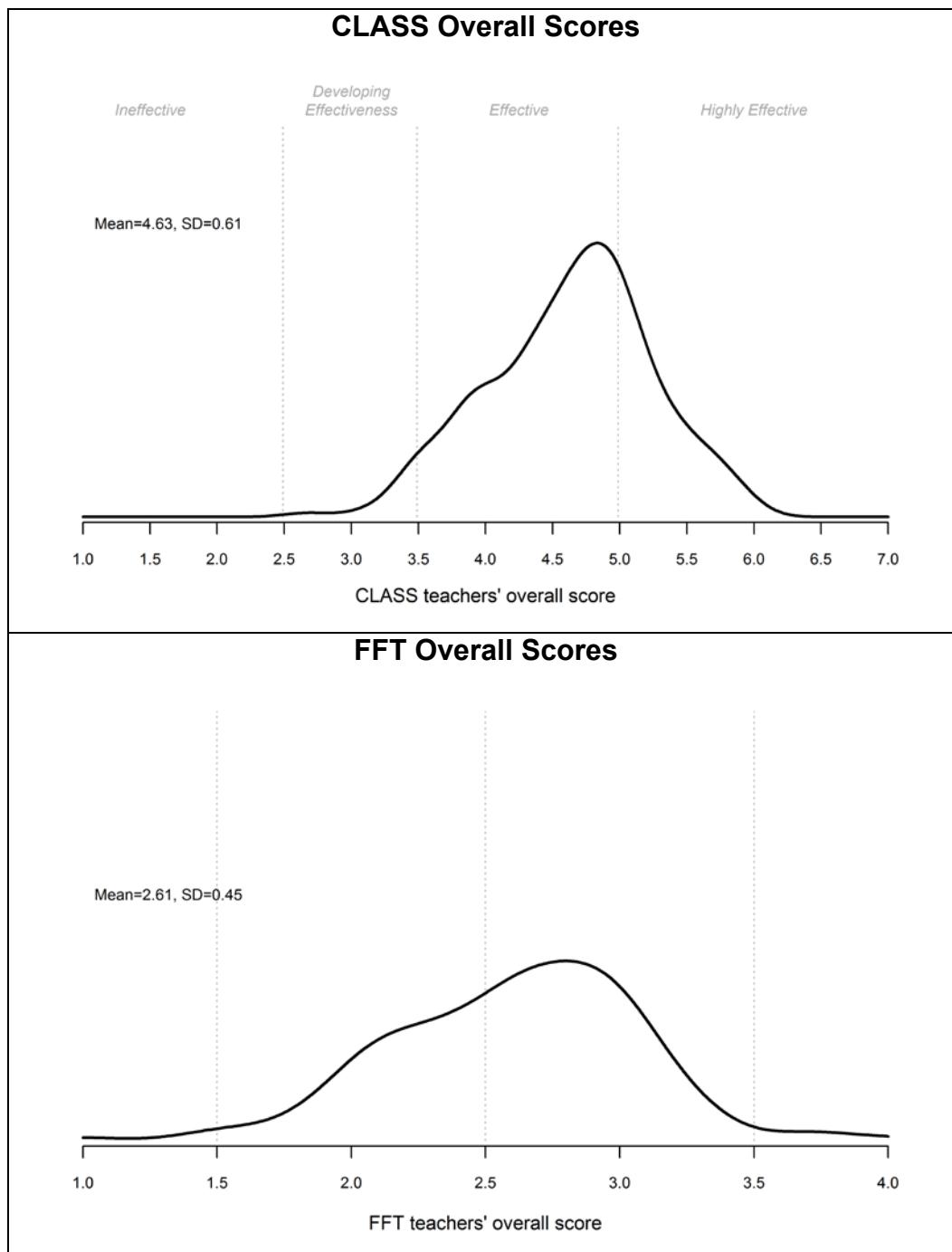
NOTES: * Difference between overall score for a specific window in Year 1 and the overall score for the same window in Year 2 is statistically significant at the .05 level (two-tailed).

† Difference between the overall score in Window 1 and the overall score in Window 4 in the same year is statistically significant at the .05 level (two-tailed).

‡ Difference between the overall score in Year 1 Window 4 and the overall score in Year 2 Window 1 is statistically significant at the .05 level (two-tailed).

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

Exhibit D.7d. Distribution of CLASS overall scores based on video-recorded lessons for treatment teachers in CLASS districts, and FFT scores for treatment teachers in FFT districts, spring Year 2



NOTES: Sample size = 238 teachers in CLASS districts and 196 teachers in FFT districts. 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 gray 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.

SOURCE: Spring 2014 Classroom Videos.

Exhibit D.7e. Pairwise correlations of intervention observation scores and video-recorded lesson scores with prior-year value-added, for treatment teachers in CLASS and FFT districts, Year 2

	Overall value-added ^a		Mathematics value-added		Reading/ELA value-added	
	N	Correlation coefficient	N	Correlation coefficient	N	Correlation coefficient
CLASS scores for treatment teachers in CLASS districts						
Intervention observation score, four-window average	249	0.18*	196	0.17*	180	0.13*
Video score, two-round average ^b	107	0.25*	87	0.27*	80	0.23*
FFT scores for treatment teachers in FFT districts						
Intervention observation score, four-window average	162	0.31*	136	0.32*	135	0.25*
Video score, two-round average ^b	92	0.10	78	0.24*	76	-0.06

NOTES: ^aThe 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.

^bThe study team rated teachers based on one video-recorded lesson in the spring and a second for a randomly selected sample of half the teachers, as explained in appendix B.

* The correlation is statistically significant at the .05 level (two-tailed).

SOURCES: Teachstone Online System (CLASS), Teachscape Online System (FFT), Spring 2014 Classroom Videos, and AIR Value-Added system.

Exhibit D.8. Descriptive statistics for average CLASS and FFT observation scores in each year, by observer type

	Score from study-hired observers			Score from school administrators			Correlation coefficient ^a
	N	Mean	Standard deviation	N	Mean	Standard deviation	
CLASS							
Year 1 overall score	294	5.80	0.58	245	5.54*	0.82	.40
Year 2 overall score	303	5.53	0.57	228	5.62	0.70	.45
FFT							
Year 1 overall score	222	3.07	0.31	221	3.04	0.38	.56
Year 2 overall score	199	3.10	0.35	193	3.03*	0.34	.49

NOTES: In cases where a teacher had more than one score from a school administrator, the scores were averaged.

* Differences between the overall score from study-hired observers and the overall score from school administrators for teachers with scores from each observer type are statistically significant at the .05 level (two-tailed).

^a Correlation coefficients are based on each teacher's mean score from study-hired observers (averaged across multiple observations) and score from the school administrator.

SOURCES: Teachstone Online System and Teachscape Online System.

Exhibit D.9a. Descriptive statistics for four-window average CLASS observation scores and two-round average video-recorded lesson scores, for treatment teachers in CLASS districts, by domain and dimension, Year 2

CLASS domains and dimensions	In-person observations (four-window average)			Video-recorded lessons (two-round average) ^a		
	N	Mean	Standard deviation	N	Mean	Standard deviation
Overall score	303	5.54	0.54	238	4.63*	0.61
Domain: Emotional support	303	5.62	0.66	238	4.44*	0.85
Positive climate	303	6.01	0.67	238	4.67*	1.03
Teacher sensitivity	303	5.77	0.70	238	4.98*	1.00
Regard for student perspectives	303	5.08	0.82	238	3.67*	1.01
Domain: Classroom organization	303	6.49	0.35	238	6.12*	0.53
Behavior management	303	6.29	0.51	238	6.06*	0.75
Productivity	303	6.25	0.53	238	5.59*	0.83
Negative climate (reverse coded)	303	6.92	0.17	238	6.71*	0.44
Domain: Instructional support	303	4.84	0.73	238	3.72*	0.78
Instructional learning formats	303	5.51	0.76	238	4.77*	0.90
Content understanding	303	5.05	0.80	238	3.92*	0.99
Analysis and inquiry	303	4.11	0.89	238	2.74*	0.94
Quality of feedback	303	4.90	0.80	238	3.84*	1.08
Instructional dialogue	303	4.65	0.91	238	3.31*	0.96
Domain: Student engagement	303	5.92	0.65	238	5.34*	0.72

NOTES: Means and standard deviations are for treatment teachers in the CLASS districts.

^aThe study team rated teachers based on one video-recorded lesson in the spring and a second for a randomly selected sample of half the teachers, as explained in appendix B. If two lessons were rated, the teacher's scores were based on an average of the two lessons; otherwise, the scores were based on the ratings from the single lesson.

* Differences between the mean in-person observation score and mean video-recorded lesson score are statistically significant at the .05 level (two-tailed).

SOURCES: Teachstone Online System and Spring 2014 Classroom Videos.

Exhibit D.9b. Descriptive statistics for four-window average FFT observation scores and two-round average video-recorded lesson scores, for treatment teachers in FFT districts, by domain and dimension, Year 2

FFT domains and dimensions	In-person observations (four-window average)			Video-recorded lessons (two-round average) ^a		
	N	Mean	Standard deviation	N	Mean	Standard deviation
Overall score	199	3.08	0.31	196	2.61*	0.45
Domain 2: Classroom environment						
Creating an environment of respect and rapport	199	3.22	0.42	196	2.86*	0.53
Establishing a culture for learning	199	3.06	0.38	196	2.66*	0.56
Managing classroom procedures	199	3.04	0.37	196	2.76*	0.49
Managing student behavior	199	3.13	0.41	196	2.92*	0.53
Organizing physical space	199	3.07	0.28	NA		
Domain 3: Instruction						
Communicating with students	199	3.26	0.41	196	2.70*	0.55
Using questioning and discussion techniques	199	2.96	0.40	196	2.19*	0.70
Engaging students in learning	199	3.02	0.43	196	2.46*	0.62
Using assessment in instruction	198	3.02	0.39	196	2.35*	0.61
Demonstrating flexibility and responsiveness	188	3.08	0.35	NA		

NOTES: Means and standard deviations are for treatment teachers in the FFT districts.

^aThe study team rated teachers based on one video-recorded lesson in the spring and a second for a randomly selected sample of half the teachers, as explained in appendix B. If two lessons were rated, the teacher's scores were based on an average of the two lessons; otherwise, the scores were based on the ratings from the single lesson.

* Differences between the mean in-person observation score and mean video-recorded lesson score are statistically significant at the .05 level (two-tailed).

NA = dimension cannot be scored with video-recorded lessons.

SOURCES: Teachscape Online System and Spring 2014 Classroom Videos.

Exhibit D.10. Percentage of teachers whose dimension scores spanned one, two, three, or four performance levels, by observation window

	Number of teachers	One level	Two levels	Three levels	Four levels
CLASS, Year 1					
Window 1	262	38.9	26.0	23.7	11.5
Window 2	307	47.9	25.1	19.5	7.5
Window 3	309	50.8	25.6	16.2	7.4
Window 4	279	57.0	20.4	17.9	4.7
CLASS, Year 2					
Window 1	297	25.9	35.7	28.0	10.4
Window 2	295	32.9	32.5	24.4	10.2
Window 3	300	45.0	29.3	21.0	4.7
Window 4	302	50.7	31.8	14.2	3.3
FFT, Year 1					
Window 1	216	31.5	62.0	6.5	0.0
Window 2	219	25.6	68.5	5.9	0.0
Window 3	220	22.7	72.7	4.6	0.0
Window 4	217	22.6	72.8	†	†
FFT, Year 2					
Window 1	191	18.3	72.3	9.4	0.0
Window 2	196	21.4	71.4	7.1	0.0
Window 3	196	21.9	70.9	7.1	0.0
Window 4	198	24.8	70.2	†	†

NOTE: † Reporting standards not met; in one or more cells, there are too few cases to report.

SOURCES: Teachstone Online System; Teachscape Online System.

Exhibit D.11. Percentage of treatment teachers who agreed somewhat or strongly with each statement about the feedback they received from the study's CLASS/FFT observations, compared with the feedback received prior to the intervention as part of their district's approach to formal evaluation, Year 2

<i>The feedback I received from the study's CLASS/FFT observations...</i>	All districts	CLASS districts	FFT districts
Was harder to understand.	30.4	31.5	29.4
Was more objective.	71.7	70.4	73.0
Was more specific about what constitutes high-quality teaching.	79.0	79.5	78.6
Was more focused on specific things I did during the observation.	84.3	80.4	88.2
Was more critical of my performance.	68.1	67.3	68.8
Provided me with clearer ideas about how my teaching could improve.	73.4	78.8	68.2
The written narratives from the study's CLASS/FFT observations provided more detail than the district's written feedback. ^a	78.3	75.9	80.6
The feedback sessions associated with the TLES observations were more useful than the district's observation feedback sessions. ^b	65.4	63.9	66.9
Number of teachers	438 or 439	267	171 or 172

NOTES: ^a Sample size for all districts = 320 teachers; Sample size for CLASS districts = 190 teachers; Sample size for FFT districts = 130 teachers.

^b Sample size for all districts = 399 teachers; Sample size for CLASS districts= 242 teachers; Sample size for FFT districts = 157 teachers.

SOURCE: Spring 2014 Teacher Survey.

Exhibit D.12. Percentage of principals who agreed somewhat or strongly with each statement about the fairness and validity of CLASS or FFT, Year 2

	All districts	CLASS districts	FFT districts
The CLASS/FFT rating system does a good job distinguishing effective from ineffective teaching.	≥ 95.0 [†]	≥ 89.0 [†]	≥ 90.0 [†]
The CLASS/FFT rating system is fair to all teachers, regardless of their personal characteristics.	92.1	≥ 89.0 [†]	≥ 90.0 [†]
The CLASS/FFT rating system is fair to all teachers, regardless of the characteristics of the students they teach.	93.7	≥ 89.0 [†]	≥ 90.0 [†]
The CLASS/FFT rating system does NOT accurately reflect the quality of an individual's teaching.	24.2	25.8	22.7
Number of principals	60	29	31

NOTE: [†] Reporting standards not met, too few cases to report the exact percentage.

SOURCE: Spring 2014 Teacher Survey.

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

Appendix E. 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.¹³⁷

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

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

where the teacher effect ($\boldsymbol{\theta}$) 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, and \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 for each teacher ($\boldsymbol{\theta}$) was generated based on the empirical Bayes estimate from the random-effects model.

¹³⁷ To account for the errors in the right-hand-side variables, we subtracted off the variance due to measurement error from the design matrix. To account for the measurement error in the left-hand-side variables, we adjusted the residual term (Doran 2014).

Covariates Included in the Models for Each District

A set of common covariates was 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.¹³⁸

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 or 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 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; 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

¹³⁸ 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.

calculating teacher overall value-added scores and school/district value-added scores, which were obtained for each year separately.

To produce an overall value-added score for each teacher for each year, we first standardized the teacher-, subject-, and 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-, and year-specific value-added scores for teachers if a teacher taught the same students in both mathematics and reading/ELA in a given year. When this happens, the covariance term would not be zero and was approximated within teacher with

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

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

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 F. Supplemental Findings About the Implementation of the Intervention’s Measure of Student Growth

Exhibit F.1. Percentage of treatment teachers with sufficient data to estimate value-added scores, and percentage of teachers whose scores were based on two years of data, by year

	Year 1	Year 2
Percentage of teachers with sufficient data to estimate value-added scores	80.1	79.8
Percentage of teachers whose scores were based on two years of data	67.7	68.0
Number of teachers	527	519

SOURCE: AIR Value-Added System.

Exhibit F.2. Percentage Distribution of treatment teachers based on whether their subject area value-added scores were measurably above or below the district average, by year

Reading value-added score	Mathematics value-added score		
	Measurably below average	Not measurably different from average	Measurably above average
Year 1 (N = 239)			
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
Year 2 (N = 235)			
Measurably below average	3.8	2.1	†
Not measurably different from average	9.8	48.9	22.1
Measurably above average	†	5.1	6.0

NOTE: † Reporting standards not met; in one or more cells, there are too few cases to report.

SOURCE: AIR Value-Added System.

Exhibit F.3. Percentage of treatment teachers who agreed somewhat or strongly with statements about the student growth reports they viewed, Year 2

Statement	All districts	CLASS districts	FFT districts
The VA (value-added) report was easy to understand.	52.5	57.3	47.8
I understand what I would need to do to improve my VA score.	54.9	57.1	52.7
The VA score is a good measure of how well students learned what I taught last year.	47.5	51.1	44.1
The VA score is a good measure of my overall performance as a teacher.	42.6	47.9	37.6
The information as an indicator of teacher effectiveness is fair to all teachers, regardless of the personal characteristics of the students they teach.	42.1	45.8	38.6
The information as an indicator of teacher effectiveness is fair to all teachers, regardless of the prior achievement of the students they teach.	40.9	46.6	35.4
Number of teachers	311–315	183–186	128 or 129

SOURCE: Spring 2014 Teacher Survey.

Exhibit F.4. Percentage of treatment principals who agreed somewhat or strongly with statements about the student growth reports they viewed, Year 2

Statement	All districts	CLASS districts	FFT districts
The information as an indicator of teacher effectiveness is fair to all teachers, regardless of the personal characteristics of the students they teach.	75.3	81.5	69.2
The information as an indicator of teacher effectiveness is fair to all teachers, regardless of the prior achievement of students they teach.	72.5	78.3	66.9
The VA score is a good measure of how well students learned what teachers in my school taught during the year.	73.7	73.7	73.7
The VA report does a good job distinguishing effective from ineffective teachers.	70.1	73.7	66.6
The VA report accurately reflects the quality of teachers who taught in this school.	72.4	82.4	62.7
Number of principals	51 or 52	26	25 or 26

SOURCE: Spring 2014 Principal Survey.

Appendix G. Supplemental Findings About the Implementation of the Intervention’s Measure of Principal Leadership

Exhibit G.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 G.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 G.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 <u>below basic</u> 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 <u>basic</u> 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 <u>proficient</u> 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 <u>distinguished</u> 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		Summary of Key Processes 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
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.

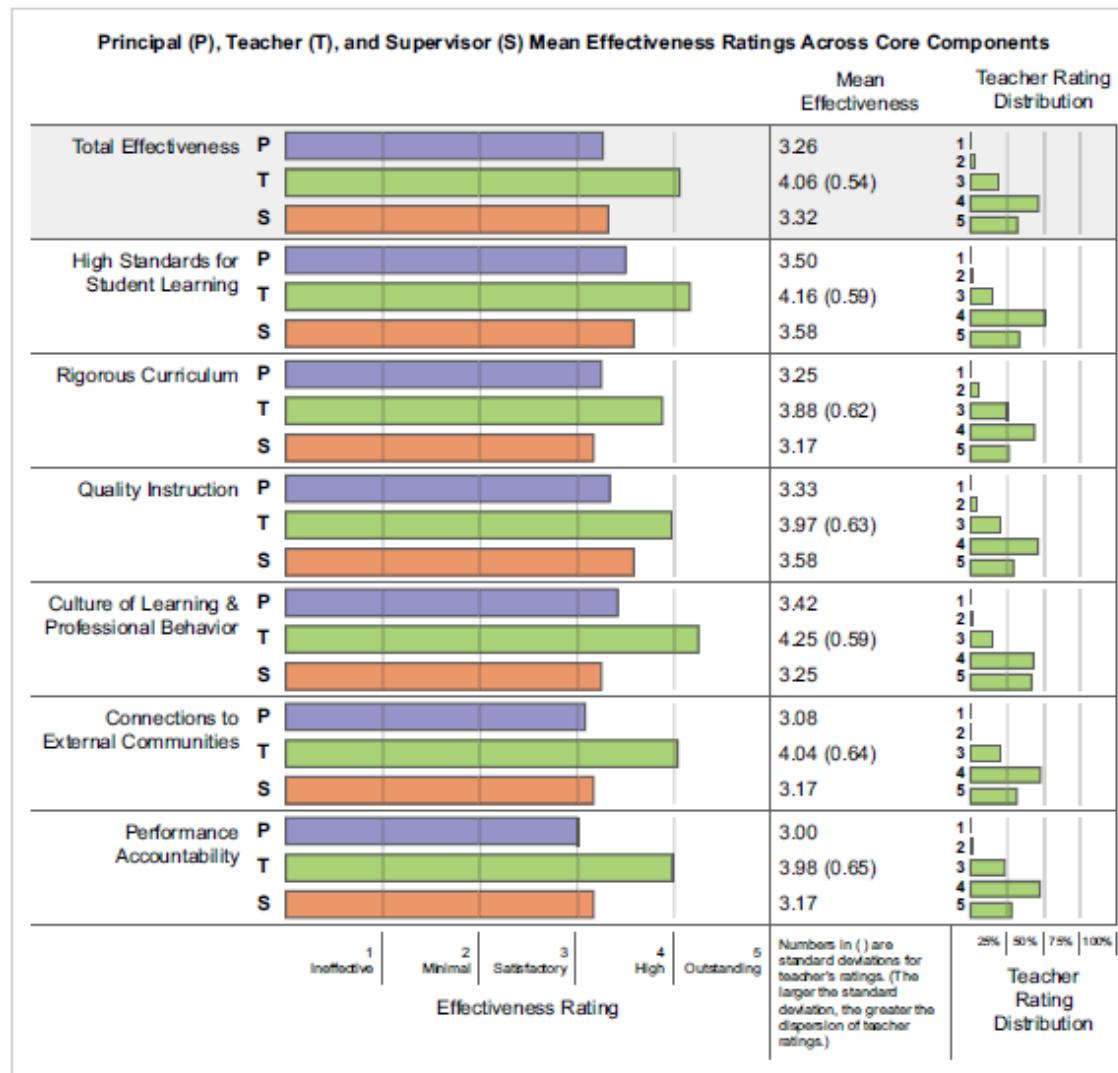
Exhibit G.4. Results by respondent group from a sample VAL-ED report

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



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 G.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 G.6. Descriptive statistics for average VAL-ED overall scores in fall and spring of each year

	Mean	Standard deviation
Year 1		
Fall overall score	3.46	0.32
Spring overall score	3.61*	0.35
Year 2		
Fall overall score	3.61	0.33
Spring overall score	3.68*†	0.35

NOTES: Sample size = 63 principals at each time point. Some principals left and were replaced by others over the two years.

* Difference between the overall score in fall and spring of the same year is statistically significant at the .05 level (two-tailed).

† Difference between the overall score in fall of Year 1 and spring of Year 2 is statistically significant at the .05 level (two-tailed).

SOURCES: Fall 2012, Spring 2013, Fall 2013, and Spring 2014 VAL-ED Surveys.

Exhibit G.7. Descriptive statistics for average VAL-ED overall scores in fall and spring of each year, by respondent group

	Score from principal		Score from supervisor		Score from teachers	
	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
Year 1						
Fall overall score	3.43	0.55	3.41	0.50	3.54	0.42
Spring overall score	3.76*†§	0.51	3.50§	0.47	3.57	0.46
Year 2						
Fall overall score	3.60	0.53	3.61	0.58	3.61	0.43
Spring overall score	3.79†§	0.50	3.70‡	0.48	3.56	0.45

NOTES: Sample size = 63 principals at each time point. Some principals left and were replaced by others over the two years.

* Difference between the rating from the principal and rating from the supervisor is statistically significant at the .05 level (two-tailed).

† Difference between the rating from the principal and rating from the teachers is statistically significant at the .05 level (two-tailed).

‡ Difference between the rating from the supervisor and rating from the teachers is statistically significant at the .05 level (two-tailed).

§ Difference between the spring overall score and fall overall score is statistically significant at the .05 level (two-tailed).

SOURCES: Fall 2012, Spring 2013, Fall 2013, and Spring 2014 VAL-ED Surveys.

Exhibit G.8. Percentage of principals whose VAL-ED scores spanned one, two, three, or four performance levels, by wave

Wave	One level	Two levels	Three levels	Four levels
Year 1				
Fall 2012	†	†	50.8	33.3
Spring 2013	†	†	60.3	28.6
Year 2				
Fall 2013	†	†	47.6	38.1
Spring 2014	†	†	44.4	36.5

NOTES: Sample size = 63 principals.

†Reporting standards not met; in one or more cells, there are too few cases to report.

SOURCES: Fall 2012, Spring 2013, Fall 2013, and Spring 2014 VAL-ED Surveys.

Exhibit G.9. Percentage of treatment principals who agreed somewhat or strongly with statements about the feedback they received from the VAL-ED, Year 2

Relative to the district's approach to evaluation...	All districts	CLASS districts	FFT districts
The feedback I received from VAL-ED observations was harder to understand.	33.5	31.9	35.0
The feedback I received from VAL-ED observations was more objective.	72.8	65.2	80.2
The feedback I received from VAL-ED observations was more specific about what constitutes high-quality leadership.	77.8	65.4	89.8
The feedback I received from VAL-ED observations was more critical of my performance.	57.8	59.8	55.8
The feedback I received from VAL-ED observations provided me with clearer ideas about how my leadership could improve.	75.0	60.7	88.8
The feedback I received from VALED observations was less comprehensive, ignoring some aspects of my role as principal.	55.4	61.0	50.0
Number of principals	45	21	24

SOURCE: Spring 2014 Principal Survey.

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

Appendix H. Technical Details About Analyses Assessing Treatment-Control Differences in Educators' Experiences and Impacts on Outcomes

This appendix provides the technical details for the following types of analyses presented in the report:

- Analyses assessing treatment-control differences in educators' experiences
- Analyses assessing the impact of the intervention on teacher outcomes
- Analyses assessing the impact of the intervention on the relationship between teachers' value-added score and their self-ratings
- Analyses assessing the impact of the intervention on principal outcomes
- Analyses assessing the impact of the intervention on student achievement
- Sensitivity analyses
- Differential impact analyses
- Analyses estimating the relationships between educator outcomes and student achievement

Analyses Assessing Treatment-Control Differences in Educators' Experiences

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. The analyses were conducted separately for each of the two study years, based on data from all teachers and principals who responded to the relevant survey questions in the spring of a given year. The specific analytic approach differed for binary survey measures (e.g., whether a teacher received ratings based on observations) and continuous survey measures (e.g., the number of instances of feedback received), as described separately below.

Analyses of Binary Measures

For binary measures of teachers' experiences with performance evaluation, we examined the treatment-control differences using a two-level linear probability model specified as follows:¹³⁹

¹³⁹ 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 teachers or 100 percent of the control 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.

Level 1 (Teachers)

$$Y_{jk} = \beta_{0k} + r_{jk} \quad (1)$$

where

- Y_{jk} is the response of teacher j in school k to a given binary survey measure;
- β_{0k} is the average response across teachers in school k ; and
- r_{jk} is a random error associated with teacher j in school k .

Level 2 (Schools)

$$\beta_{0k} = \sum_{b=1}^{37} \gamma_{00b} B_{bk} + \sum_{d=1}^8 \gamma_{01d} (T * D_d)_k + u_{0k} \quad (2)$$

where

- 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_{0k} is a random error associated with school k .

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 eight 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 of 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.

Analyses Assessing the Impact of the Intervention on Teacher Outcomes

In this section, we describe the analytic models used to assess the impact of the intervention on teachers' initial outcomes and teachers' classroom practice. Teachers' initial outcomes (e.g., their self-ratings and their interest in improving specific areas of practice) were measured with a teacher survey administered in the spring of each of the two study years. Separate analyses of teachers' initial outcomes were conducted for each year, based on data from all teachers who responded to the relevant survey questions in the spring of a given year. Classroom practice was measured only in the spring of the second year based on classroom videos collected from grade 4–8 reading/ELA and mathematics teachers who were present in study schools in the spring of the second year.

Analyses of Impact on Teachers' Initial Outcomes

One initial outcome that we examined is teachers' ratings of their own performance in improving student achievement relative to other teachers in their district. The rating categories on the teach survey were: 1 = Very Poor (bottom 5%); 2 = Poor (6th–25th percentile); 3 = Fair (26th–50th percentile); 4 = Good (51st–75th percentile); 5 = Very Good (76th–95th percentile); and 6 = Exceptional (top 5%). To facilitate the interpretation of findings, we converted the rating categories to a percentile scale by replacing each rating category with the midpoint in the percentile range corresponding to that category (i.e., recoding the six categories into 3, 15.5, 38, 63, 85.5, and 98, respectively). The model for assessing the impact of the intervention on teachers' self-ratings using the percentile scale is specified as follows:

Level 1 (Teachers)

$$Y_{jk} = \beta_{0k} + \sum_{p=1}^4 \beta_{1kp} W_{pjk} + r_{jk} \quad (3)$$

where

- Y_{jk} is the self-rating in percentile scale for teacher j in school k ;
- W_{pjk} , $p = 1, 2, 3$, and 4 , is a vector of background characteristics for teacher j in school k , including three dummy indicators for years of teaching experience (i.e., 4–10 years, 11–20 years, and more than 20 years, with 3 or fewer years as the omitted reference) and one dummy indicator for whether the teacher had a master's degree or higher, grand-mean centered;

- β_{0k} is the average self-rating of teachers in school k , adjusted for teacher background characteristics;
- β_{1kp} represents the relationship between teacher background characteristic p and teachers' self-ratings in school k ; and
- r_{jk} is a random error associated with teacher j in school k .

Level 2 (Schools)

$$\beta_{0k} = \sum_{b=1}^{37} \gamma_{00b} B_{bk} + \sum_{d=1}^8 \gamma_{01d} (T^* D_d)_k + u_{0k} \quad (4)$$

$$\beta_{1kp} = \gamma_{10p} \quad (5)$$

where

- B_{bk} and $(T^* D_d)_k$ are defined as in equation 2;
- γ_{00b} is the adjusted average self-rating of teachers in control schools in block b ;
- γ_{01d} is the difference between treatment and control schools in teachers' self-ratings in district d ;
- γ_{10p} is the average relationship between teacher background characteristic p and teachers' self-ratings across all schools; and
- u_{0k} is a random error associated with school k .

The estimate of primary interest from the above model is γ_{01d} , $d = 1–8$, which represents the impact of the intervention on teachers' self-ratings within each district. The average impact across all eight districts was computed as a weighted average, with each district weighted by the number of treatment schools in the district. This weighted overall impact represents the impact for a typical treatment school in the study sample.

The intervention's impact on teachers' initial outcomes based on binary survey measures (e.g., whether a teacher was interested in improving in a CLASS/FFT-related area) was assessed using a two-level linear probability model specified in the same way as the model described above.

Analyses of Impact on Classroom Practice

To assess the intervention's impact on teachers' classroom practice, we coded all of the video-recorded lessons from the spring of the second year using both CLASS and FFT. Because about half of the teachers were observed twice, we estimated the intervention's impact on classroom

practice using the following three-level model, which explicitly takes into account the clustering of lessons within teachers and teachers within schools:

Level 1 (Lessons)

$$Y_{ijk} = \pi_{0jk} + \varepsilon_{ijk} \quad (6)$$

where

- Y_{ijk} is a measure of classroom practice (e.g., the CLASS or FFT overall score) for lesson i taught by teacher j from school k ;
- π_{0jk} is the average classroom practice score across lessons for teacher j from school k ; and
- ε_{ijk} is a random error associated with lesson i taught by teacher j from school k .

Level 2 (Teachers)

$$\pi_{0jk} = \beta_{00k} + \sum_{p=1}^4 \beta_{01kp} W_{pj} + r_{0jk} \quad (7)$$

where

- W_{pj} is defined as in equation 3;
- β_{00k} is the average classroom practice score across teachers in school k , adjusted for teacher background characteristics;
- β_{01kp} is the relationship between teacher background characteristic p and teachers' classroom practice scores in school k ; and
- r_{0jk} is a random error associated with teacher j in school k .

Level 3 (Schools)

$$\beta_{00k} = \sum_{b=1}^{37} \gamma_{000b} B_{bk} + \sum_{d=1}^8 \gamma_{001d} (T^* D_d)_k + u_{00k} \quad (8)$$

$$\beta_{01kp} = \gamma_{010p} \quad (9)$$

where

- B_{bk} and $(T^* D_d)_k$ are defined as in equation 2;
- γ_{000b} is the adjusted average classroom practice score of teachers in control schools in block b ;
- γ_{001d} is the difference between treatment and control schools in teachers' classroom practice scores in district d ;

- γ_{010_p} is the average relationship between teacher background characteristic p and classroom practice scores across all schools; and
- u_{00k} is a random error associated with school k .

The estimate of primary interest from the above model is γ_{001d} , $d = 1-8$, which represents the impact of the intervention on classroom practice within each district. The average impact across all eight districts was computed as a weighted average, with each district weighted by the number of treatment schools in the district. Separate average impacts were also computed across the four CLASS districts and across the four FFT districts, and the difference in impact between the CLASS districts and FFT districts was tested using a Z test.

Handling of Missing Data

Teachers with missing outcome data were excluded from the impact analyses described above. Missing covariate data were handled using the dummy variable adjustment approach (Puma et al. 2009). For each teacher-level covariate with missing data, we set the missing value to zero and included a missingness indicator in the impact model. Missing data were handled using the same approach for analyses of the intervention's impact on principal outcomes and student outcomes described in later sections. For simplicity, the missingness indicators are not shown in the impact models presented in this appendix.

Analyses Assessing the Impact of the Intervention on the Relationship Between Teachers' Value-Added Scores and Their Self-Ratings

In addition to assessing whether teachers' self-ratings were higher in treatment schools than in control schools, we also examined whether teachers' self-ratings were more strongly correlated with their prior value-added scores in treatment schools than in control schools. This analysis was based on the following model:

Level 1 (Teachers)

$$Y_{jk} = \beta_{0k} + \sum_{p=1}^4 \beta_{1kp} W_{pjk} + \beta_{2k} PRIOR_VA_{jk} + r_{jk} \quad (10)$$

where

- W_{pjk} , β_{1kp} , and r_{jk} are defined as in equation 3;
- Y_{jk} is the self-rating in percentile scale for teacher j in school k ;
- $PRIOR_VA_{jk}$ is the percentile rank of the prior value-added score for teacher j in school k ;
- β_{0k} is the average self-rating of teachers in school k , adjusted for teacher background characteristics and prior value-added scores; and

- β_{2k} represents the relationship between teachers' prior value-added scores and their self-ratings in school k .

Level 2 (Schools)

$$\beta_{0k} = \sum_{b=1}^{37} \gamma_{00b} B_{bk} + \gamma_{01} T_k + u_{0k} \quad (11)$$

$$\beta_{1kp} = \gamma_{10p} \quad (12)$$

$$\beta_{2k} = \sum_{b=1}^{37} \gamma_{20b} B_{bk} + \gamma_{21} T_k \quad (13)$$

where

- B_{bk} is defined as in equation 2, γ_{00b} and u_{0k} are defined as in equation 4, and γ_{10p} is defined as in equation 5;
- T_k is a dummy indicator for treatment status, coded 1 for treatment schools and 0 for control schools;
- γ_{20b} represents the relationship between teachers' prior value-added scores and their self-ratings in control schools in block b ; and
- γ_{21} represents the difference between treatment and control schools in the relationship between teachers' prior value-added scores and their self-ratings.

The estimate of primary interest from the above model is γ_{21} . A positive value of γ_{21} would indicate that a teacher's prior value-added score was more strongly correlated with the teacher's self-rating in treatment schools than in control schools.

Analyses Assessing the Impact of the Intervention on Principal Outcomes

This section presents the analytic models used to assess the impact of the intervention on principals' initial outcomes and principal leadership. Principals' initial outcomes (e.g., their self-ratings and their interest in improving specific areas of practice) were measured with a principal survey administered in the spring of each of the two years of intervention. Principal leadership was measured with a teacher survey administered in the spring of both years. Separate analyses of principal outcomes were conducted for each year, based on data from all teachers and principals who responded to the relevant survey questions in the spring of a given year.

Analyses of Impact on Principals' Initial Outcomes

We estimated the intervention's impact on principals' self-ratings using a principal-level regression as specified below. Impact on initial outcomes based on binary principal survey measures (e.g., whether the principal wanted to improve in a VAL-ED-related area) was assessed using a linear probability model specified similarly.

$$Y_k = \sum_{b=1}^{37} \gamma_{0b} B_{bk} + \sum_{d=1}^8 \gamma_{1d} (T^* D_d)_k + \sum_{q=1}^3 \gamma_{2q} Z_{qk} + u_k \quad (14)$$

where

- B_{bk} and $(T^* D_d)_k$ are defined as in equation 2;
- Y_k is the self-rating (in the percentile scale) of principal k ;
- Z_{qk} , $q = 1, 2$, and 3 , is a vector of principal background characteristics, including two dummy indicators for years of experience as a principal (i.e., 4–10 years and more than 10 years, with 3 or fewer years as the omitted reference) and a continuous variable for years of teaching experience;
- γ_{0b} is the average self-rating of control principals in block b , adjusted for principal background characteristics;
- γ_{1d} is the difference between treatment and control principals in their self-ratings in district d ;
- γ_{2q} is the relationship between principal background characteristic q and principals' self-ratings; 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 impact of the intervention on principals' self-ratings in each study district. The average impact across all eight districts was computed as a weighted average, with each district weighted by the number of treatment schools in the district.

Analyses of Impact on Principal Leadership

We estimated the intervention's impact on principal leadership using two scales created based on data from the spring teacher surveys: *principal instructional leadership* and *teacher-principal trust*. Because the two principal leadership scales are teacher-level measures, the analyses were conducted using the same model that was used to assess impact on teachers' self-ratings (see equations 3–5). In addition to overall impact across all study districts, we estimated separate impacts for the CLASS districts and the FFT districts, and we tested the difference in impact between the two sets of districts using a Z test.

Analyses Assessing the Impact of the Intervention on Student Achievement

To estimate the impact of the intervention on student achievement, we used a three-level model (where students were nested within teachers and teachers nested within schools) with data pooled from grades 4–8 across the eight study districts. The impact analyses were conducted separately

by year (Year 1 and Year 2) and by subject (mathematics and reading/ELA), and included grade 4–8 students who were in the reading/ELA or mathematics classes taught by study teachers in the spring of a given year. The model is specified as follows:

Level 1 (Students)

$$Y_{ijk} = \pi_{0jk} + \pi_{1jk} P_{ijk} + \pi_{2jk} GRD_AVG_{ijk} + \sum_{g=5}^8 \pi_{3jkg} G_{gijk} + \sum_{m=1}^5 \pi_{4jkm} X_{mijk} + \varepsilon_{ijk} \quad (15)$$

where

- Y_{ijk} is the standardized test score for student i taught by teacher j in school k ;
- P_{ijk} is the test score from the baseline year for student i taught by teacher j in school k , grand-mean centered;¹⁴⁰
- GRD_AVG_{ijk} is the school average test score from the baseline year for the grade the student was in for the impact analysis;¹⁴¹
- G_{gijk} , $g = 5, 6, 7$, and 8 , is a set of grade indicators for each student, with grade 4 being the omitted reference grade, grand-mean centered;
- X_{mijk} , $m = 1, 2, \dots, 5$, is a vector of demographic characteristics for each student, including gender, race (White versus non-White), eligibility for free or reduced-price lunch, English learner status, and special education status, grand-mean centered;
- π_{0jk} is the average test score among students taught by teacher j in school k , adjusted for student characteristics;
- π_{1jk} , π_{2jk} , π_{3jkg} , and π_{4jkm} represent the relationships between the student characteristics included in the model and their test scores among students taught by teacher j in school k , each adjusted for the other variables in the model; and
- ε_{ijk} is a random error associated with student i taught by teacher j in school k .

Level 2 (Teachers)

$$\pi_{0jk} = \beta_{00k} + r_{0jk} \quad (16)$$

$$\pi_{1jk} = \beta_{10k} \quad (17)$$

$$\pi_{2jk} = \beta_{20k} \quad (18)$$

¹⁴⁰ For the first-year impact analysis, this variable represents a student’s test score from the prior/baseline year. For the second-year impact analysis, this variable represents a student’s test score from two years prior and is thus a less strong predictor than in the first-year analysis.

¹⁴¹ For a grade 4 student in the first-year impact analysis, for example, this variable represents the school average test score for grade 4 students in the prior/baseline year. For the second-year impact analysis, this variable represents the grade-specific school average test score from two years prior and is thus a less strong predictor than in the first-year analysis.

$$\pi_{3jkg} = \beta_{30kg}, g = 5-8 \quad (19)$$

$$\pi_{4jkm} = \beta_{40km}, m = 1-5 \quad (20)$$

where

- β_{00k} is the adjusted classroom average test score across all teachers in school k ;
- β_{10k} , β_{20k} , β_{30kg} , and β_{40km} represent the adjusted average relationships between the student characteristics and their test scores across all classrooms taught by teachers in school k ; and
- r_{0jk} is a random error associated with teacher j in school k .

Level 3 (Schools)

$$\beta_{00k} = \sum_{b=1}^{37} \gamma_{000b} B_{bk} + \sum_{d=1}^8 \gamma_{001d} (T^* D_d)_k + u_{00k} \quad (21)$$

$$\beta_{10k} = \sum_{d=1}^8 \gamma_{100d} D_{dk} \quad (22)$$

$$\beta_{20k} = \sum_{d=1}^8 \gamma_{200d} D_{dk} \quad (23)$$

$$\beta_{30kg} = \gamma_{300g}, g = 5-8 \quad (24)$$

$$\beta_{40km} = \gamma_{400m}, m = 1-5 \quad (25)$$

where

- B_{bk} and $(T^* D_d)_k$ are defined as in equation 2;
- D_{dk} , $d = 1-8$, is a set of dummy indicators for the eight study districts;
- γ_{000b} represents the adjusted average test score in control schools in block b ;
- γ_{001d} is the difference between treatment and control schools in the average test score;
- γ_{100d} , γ_{200d} , γ_{300g} , and γ_{400m} represent the adjusted average relationships between the student characteristics and their test scores across all schools; and
- u_{00k} is a random error associated with school k .

The estimate of primary interest from the above model is γ_{001d} , $d = 1-8$, which represents the impact of the intervention on student achievement in each of the eight study districts. The average impact across all eight districts was computed as a weighted average; each district was weighted by the number of treatment schools in the district. Separate average impacts were also

computed across the four CLASS districts and across the four FFT districts, and the difference in impact between the CLASS districts and FFT districts was tested using a Z test.

Sensitivity Analyses

For outcomes that were the primary focus of the study (i.e., classroom practice, principal leadership, and student achievement), we conducted supplemental analyses to test the sensitivity of the impact findings to alternative model specification and sample definition. First, for all three sets of outcomes, we estimated impact models that included random assignment blocks, but no covariates. The impact analyses without covariates made fewer assumptions and were expected to produce similar point estimates but larger standard errors for the treatment effects relative to impact analyses with covariates for randomized controlled trials. Second, for the classroom practice outcomes, we estimated a model that excluded one district, which had a much lower response rate on video observations than the other districts. Third, for student achievement outcomes, we estimated a model that included prior achievement scores in both reading/ELA and mathematics as separate covariates, which was expected to produce more precise impact estimates than the main impact model, where only prior achievement score in the same subject as the outcome was included as a covariate.

Differential Impact Analyses

For primary outcomes of the study, in addition to overall impact, we also explored whether the impact varied by school level and by teachers' probationary status and prior value-added score, as described below.

Analyses of Differential Impact on Classroom Practice

To test for the differential impact of the intervention on classroom practice, we modified the main impact model presented in equations 6–9 by incorporating cross-level interactions between teachers' probationary status ($PROBATION_{jk}$, coded 1 if a teacher was a probationary teacher and 0 if a nonprobationary teacher) and the district-specific treatment indicators ($(T^*D_d)_k$). The modified model is specified as follows:

Level 1 (Lessons)

$$Y_{ijk} = \pi_{0jk} + \varepsilon_{ijk} \quad (26)$$

Level 2 (Teachers)

$$\pi_{0jk} = \beta_{00k} + \sum_{p=1}^4 \beta_{01kp} W_{pj} + \beta_{02k} PROBATION_{jk} + r_{0jk} \quad (27)$$

Level 3 (Schools)

$$\beta_{00k} = \sum_{b=1}^{37} \gamma_{000b} B_{bk} + \sum_{d=1}^8 \gamma_{001d} (T^*D_d)_k + u_{00k} \quad (28)$$

$$\beta_{01kp} = \gamma_{010p} \quad (29)$$

$$\beta_{02k} = \sum_{d=1}^8 \gamma_{020d} (T * D_d)_k \quad (30)$$

The estimate of primary interest from the above model is γ_{020d} , $d = 1–8$, which represents the difference between impact for probationary teachers and impact for nonprobationary teachers within each district. The average differential impact across all eight districts was computed as a weighted average, with each district weighted by the number of treatment schools in the district.

We estimated the differential impact based on teachers' prior value-added score using the same model described above, except that the indicator for probationary status in equation 27) was replaced with a teacher's overall value-added score produced for the Wave 1 student growth reports.¹⁴² The resulting estimate captures the difference in impact between teachers whose prior value-added scores differed by one standard deviation (in student-level standard deviation units).

To estimate the differential impact on classroom practice by school level, we modified the main impact model presented in equations 6–9 by adding to equation 8 a set of interactions between a school-level indicator ($MIDDLE_k$, coded 1 for middle school and 0 for elementary school) and the district-specific treatment indicators $((T * D_d)_k)$. The expanded equation is specified as follows:

$$\beta_{00k} = \sum_{b=1}^{37} \gamma_{000b} B_{bk} + \sum_{d=1}^8 \gamma_{001d} (T * D_d)_k + \sum_{d=1}^8 \gamma_{002d} (T * D_d * MIDDLE)_k + u_{00k} \quad (31)$$

The coefficients for the three-way interactions from the above equation, γ_{002d} , $d = 1–8$, represent the difference between impact for elementary schools and impact for middle schools within each district. The average differential impact across all eight districts was computed as a weighted average; each district was weighted by the number of treatment schools in the district.

Analyses of Differential Impact on Principal Leadership

For principal leadership outcomes (instructional leadership and teacher-principal trust), we examined whether the impact of the intervention differed between elementary schools and middle schools. To do so, we modified the main impact model for principal outcomes (equation 14) by adding a school level indicator ($MIDDLE_k$) and a set of treatment-by-district-by-school-level interactions $(T * D_d * MIDDLE)_k$, as shown below:

¹⁴² The value-added scores were computed for all grade 4–8 reading/ELA and mathematics teachers with the relevant data in each study district, although the student growth reports as part of the intervention were only produced for treatment teachers.

$$Y_k = \sum_{b=1}^{37} \gamma_{0b} B_{bk} + \sum_{d=1}^8 \gamma_{1d} (T * D_d)_k + \sum_{q=1}^3 \gamma_{2q} Z_{qk} + \gamma_3 MIDDLE_k + \sum_{d=1}^8 \gamma_{4d} (T * D_d * MIDDLE)_k + u_k \quad (32)$$

The coefficients for the three-way interactions from the above equation, γ_{4d} , $d = 1–8$, represent the difference between impact for elementary schools and impact for middle schools within each district. The average differential impact across all eight districts was computed as a weighted average, with each district weighted by the number of treatment schools in the district.

Analyses of Differential Impact on Student Achievement

For student achievement, we examined whether the impact of the intervention varied by school level and by teachers' probationary status and prior value-added score. This set of differential impact analyses were conducted by modifying the teacher-level and school-level equations in the main student achievement impact model (see equations 15–25) in ways similar to how we estimated differential impact on classroom practice outcomes as described earlier.

Analyses Estimating the Relationships Between Educator Outcomes and Student Achievement

In addition to the impact analyses described above, we conducted a set of correlational analyses to examine the relationships between key educator outcomes (i.e., classroom practice and principal leadership) and student achievement, which are described in this section.

Analyses of the Relationship Between Classroom Practice and Student Achievement

Given that classroom practice for teachers in both treatment and control schools was measured only in the second year, we examined the relationship between classroom practice—as measured by the CLASS and FFT overall scores—and student achievement only in the second year. The analyses were conducted separately for mathematics and reading/ELA based on a modified achievement impact model, where we added a classroom practice measure as a predictor to equation 16 at the teacher level and fixed the coefficient for the predictor to its grand mean at the school level. The school-level coefficient for the classroom practice measure would then represent the relationship between classroom practice and student achievement in the second year, adjusted for the covariates included in the model.

Analyses of the Relationship Between Principal Leadership and Student Achievement

Our primary measures of principal leadership are two scales (*principal instructional leadership* and *teacher-principal trust*) created based on teacher surveys administered in both years. To examine the relationship between principal leadership and student achievement, we added a given principal leadership measure as a predictor to the student achievement impact model, and estimated the relationship separately for each principal leadership scale, each subject, and each year. Although the principal leadership scales were measured at the teacher level, we treated

them as school-level measures in the correlational analyses because the relevant teacher survey items were designed to tap school- and principal-level conditions. We created school-level measures of principal leadership by aggregating the teacher-level principal leadership scales to the school level, and each school-level principal leadership measure was then added to equation 21 in the student achievement impact model. The coefficient for the school-level principal leadership measure would then represent the relationship between principal leadership and student achievement, adjusted for the covariates included in the model.

Appendix I. Supporting Exhibits for Analyses of Educators' Experiences and Initial Outcomes

Supporting Exhibits for Analyses of the Performance Feedback Teachers and Principals Received

Exhibit I.1a. Percentage of teachers who reported receiving ratings on their classroom practice, being observed by their principal, and being observed by someone from outside of their school, overall and within CLASS and FFT districts, by treatment status, Year 1

	Treatment group mean	Control group mean	Estimated difference	p value
Grade 4–8 teachers in all districts				
Received rating	83.5	38.6	44.9*	0.000
Nonprobationary teachers	83.6	31.3	52.4*	0.000
Probationary teachers	84.1	68.8	15.3*	0.037
Observed by principal	83.7	76.4	7.3*	0.014
Observed by someone from outside of their school	75.0	16.1	58.9*	0.000
Grade 4–8 teachers in CLASS districts				
Received rating	78.4	37.4	41.0*	0.000
Nonprobationary teachers	77.7	32.0	45.6*	0.000
Probationary teachers	83.3	61.6	21.7*	0.033
Observed by principal	78.4	72.4	6.1*	0.158
Observed by someone from outside of their school	68.2	13.2	55.0*	0.000
Grade 4–8 teachers in FFT districts				
Received rating	88.4	39.8	48.6*	0.000
Nonprobationary teachers	92.3	33.4	58.9*	0.000
Probationary teachers	87.6	78.5	9.2	0.403
Observed by principal	88.7	80.2	8.5*	0.029
Observed by someone from outside of their school	81.5	19.0	62.5*	0.000
Grade K–3 teachers in all districts				
Received rating	77.7	40.1	37.6*	0.000
Nonprobationary teachers	80.0	35.5	44.5*	0.000
Probationary teachers	77.6	82.6	-5.0	0.587
Observed by principal	88.2	80.7	7.5*	0.007
Observed by someone from outside of their school	61.7	18.3	43.4	0.000

NOTES: Sample size for grade 4–8 teachers in all districts = 63 schools and 93–523 teachers for the treatment group; 64 schools and 120–549 teachers for the control group. Sample size for grade 4–8 teachers in CLASS districts = 31 schools and 39–305 teachers for the treatment group; 32 schools and 72–324 teachers for the control group. Sample size for grade 4–8 teachers in FFT districts = 32 schools and 54–218 teachers for the treatment group; 32 schools and 48–225 teachers for the control group. Sample size for grade K–3 teachers in all districts = 50 schools and 75–523 teachers for the treatment group; 50 schools and 74–549 teachers for the control group.

The analyses were based on a two-level analysis (teachers within schools) controlling for random assignment blocks.

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

SOURCE: Spring 2013 Teacher Survey.

Exhibit I.1b. Percentage of teachers who reported receiving ratings on their classroom practice, being observed by their principal, and being observed by someone from outside of their school, overall and within CLASS and FFT districts, by treatment status, Year 2

	Treatment group mean	Control group mean	Estimated difference	p value
Grade 4–8 teachers in all districts				
Received rating	85.6	43.1	42.6*	0.000
Nonprobationary teachers	86.5	34.9	51.7*	0.000
Probationary teachers	82.4	59.9	22.5*	0.000
Observed by principal	74.5	74.8	-0.3	0.946
Observed by someone from outside of their school	86.6	14.7	71.9*	0.000
Grade 4–8 teachers in CLASS districts				
Received rating	84.8	45.1	39.7*	0.000
Nonprobationary teachers	86.3	42.0	44.3*	0.000
Probationary teachers	80.7	53.8	26.9*	0.001
Observed by principal	66.6	66.5	0.1	0.991
Observed by someone from outside of their school	86.6	11.2	75.4*	0.000
Grade 4–8 teachers in FFT districts				
Received rating	86.5	41.0	45.4*	0.000
Nonprobationary teachers	86.8	28.3	58.5*	0.000
Probationary teachers	84.1	66.1	18.0*	0.025
Observed by principal	82.3	82.9	-0.6	0.895
Observed by someone from outside of their school	86.5	18.3	68.3*	0.000
Grade K–3 teachers in all districts				
Received rating	86.1	41.7	44.3*	0.000
Nonprobationary teachers	87.2	36.1	51.1*	0.000
Probationary teachers	84.0	51.6	32.4*	0.000
Observed by principal	78.3	76.2	2.1	0.606
Observed by someone from outside of their school	75.2	15.6	59.6*	0.000

NOTES: Sample size for grade 4–8 teachers in all districts = 63 schools and 145–495 teachers for the treatment group; 63 schools and 197–521 teachers for the control group. Sample size for grade 4–8 teachers in CLASS districts = 31 schools and 80–297 teachers for the treatment group; 32 schools and 113–310 teachers for the control group. Sample size for grade 4–8 teachers in FFT districts = 32 schools and 65–198 teachers for the treatment group; 31 schools and 84–211 teachers for the control group. Sample size for grade K–3 teachers in all districts = 49 schools and 217–662 teachers for the treatment group; 47 schools and 201–656 teachers for the control group.

The analyses were based on a two-level analysis (teachers within schools) controlling for random assignment blocks.

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

SOURCE: Spring 2014 Teacher Survey.

Exhibit I.2a. Number of feedback instances and duration of feedback that an average teacher reported receiving, overall and within CLASS and FFT districts, by treatment status, Year 1

	Treatment group median	Control group median	Estimated difference	p value
Grade 4–8 teachers in all districts				
Number of instances with any type of feedback	4.0	3.1	0.9*	0.000
Number of feedback sessions with ratings and written narrative	3.0	0.7	2.3*	0.000
Total length of oral feedback	80.0	17.9	62.1*	0.000
Grade 4–8 teachers in CLASS districts				
Number of instances with any type of feedback	4.0	3.0	1.0*	0.000
Number of feedback sessions with ratings and written narrative	3.0	1.0	2.0*	0.000
Total length of oral feedback	60.0	6.5	53.5*	0.000
Grade 4–8 teachers in FFT districts				
Number of instances with any type of feedback	4.0	3.3	0.7*	0.000
Number of feedback sessions with ratings and written narrative	3.0	0.2	2.8*	0.000
Total length of oral feedback	95.0	19.4	75.6*	0.000
Grade K–3 teachers in all districts				
Number of instances with any type of feedback	2.0	2.0	0.0*	0.934
Number of feedback sessions with ratings and written narrative	1.0	0.1	0.9*	0.000
Total length of oral feedback	45.0	17.8	27.2*	0.000

NOTES: Sample size for grade 4–8 teachers in all districts = 63 schools and 523 teachers for the treatment group; 64 schools and 549 teachers for the control group. Sample size for grade 4–8 teachers in CLASS districts = 31 schools and 305 teachers for the treatment group; 32 schools and 324 teachers for the control group. Sample size for grade 4–8 teachers in FFT districts = 32 schools and 218 teachers for the treatment group; 32 schools and 225 teachers for the control group. Sample size for grade K–3 teachers in all districts = 50 schools and 635 teachers for the treatment group; 50 schools and 664 teachers for the control group. The analyses were based on an aligned rank sum test with randomization inference about median difference between treatment and control groups.

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

SOURCE: Spring 2013 Teacher Survey.

Exhibit I.2b. Number of feedback instances and duration of feedback that an average teacher reported receiving, overall and within CLASS and FFT districts, by treatment status, Year 2

	Treatment group median	Control group median	Estimated difference	p value
Grade 4–8 teachers in all districts				
Number of instances with any type of feedback	4.0	2.8	1.2*	0.000
Number of feedback sessions with ratings and written narrative	3.0	0.2	2.8*	0.000
Total length of oral feedback	100.0	25.0	75.0*	0.000
Grade 4–8 teachers in CLASS districts				
Number of instances with any type of feedback	4.0	2.8	1.2*	0.000
Number of feedback sessions with ratings and written narrative	3.0	0.2	2.8*	0.000
Total length of oral feedback	90.0	20.2	69.8*	0.000
Grade 4–8 teachers in FFT districts				
Number of instances with any type of feedback	4.0	3.0	1.0*	0.028
Number of feedback sessions with ratings and written narrative	3.0	0.2	2.8*	0.000
Total length of oral feedback	120.0	36.3	83.7*	0.000
Grade K–3 teachers in all districts				
Number of instances with any type of feedback	2.0	2.2	-0.2	0.330
Number of feedback sessions with ratings and written narrative	2.0	0.9	1.1*	0.000
Total length of oral feedback	55.0	26.8	28.2*	0.000

NOTES: Sample size for grade 4–8 teachers in all districts = 63 schools and 495 teachers for the treatment group; 63 schools and 521 teachers for the control group. Sample size for grade 4–8 teachers in CLASS districts = 31 schools and 297 teachers for the treatment group; 32 schools and 310 teachers for the control group. Sample size for grade 4–8 teachers in FFT districts = 32 schools and 198 teachers for the treatment group; 31 schools and 211 teachers for the control group. Sample size for grade K–3 teachers in all districts = 49 schools and 662 teachers for the treatment group; 47 schools and 656 teachers for the control group. The analyses were based on an aligned rank-sum test with randomization inference about median difference between treatment and control groups.

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

SOURCE: Spring 2014 Teacher Survey.

Exhibit I.3a. Percentage of teachers who reported receiving specific types of student achievement information, overall and within CLASS and FFT districts, by treatment status, Year 1

	Treatment group mean	Control group mean	Estimated difference	p value
Grade 4–8 teachers in all districts				
Value-added scores for me based upon the students that I taught	44.7	24.2	20.5*	0.000
Data on individual students that I taught	63.7	83.9	-20.3*	0.000
Average data for classes of students that I taught	51.2	62.4	-11.2*	0.000
I did not receive any student achievement information based on standardized test results.	15.5	6.5	9.0*	0.000
Grade 4–8 teachers in CLASS districts				
Value-added scores for me based upon the students that I taught	37.7	29.4	8.4	0.059
Data on individual students that I taught	63.3	85.4	-22.1*	0.000
Average data for classes of students that I taught	49.3	64.3	-15.0*	0.000
I did not receive any student achievement information based on standardized test results.	15.9	4.9	11.0*	0.000
Grade 4–8 teachers in FFT districts				
Value-added scores for me based upon the students that I taught	51.5	19.5	31.9*	0.000
Data on individual students that I taught	64.1	82.6	-18.5*	0.000
Average data for classes of students that I taught	53.1	60.7	-7.6	0.101
I did not receive any student achievement information based on standardized test results.	15.0	8.1	6.9	0.056
Grade K–3 teachers in all districts				
Value-added scores for me based upon the students that I taught	16.5	19.3	-2.8	0.269
Data on individual students that I taught	60.0	73.4	-13.4*	0.000
Average data for classes of students that I taught	43.7	54.0	-10.3*	0.002
I did not receive any student achievement information based on standardized test results.	31.3	16.8	14.5*	0.000

NOTES: Sample size for grade 4–8 teachers in all districts = 63 schools and 519 teachers for the treatment group; 64 schools and 554 teachers for the control group. Sample size for grade 4–8 teachers in CLASS districts = 31 schools and 302 teachers for the treatment group; 32 schools and 326 teachers for the control group. Sample size for grade 4–8 teachers in FFT districts = 32 schools and 217 teachers for the treatment group; 32 schools and 228 teachers for the control group. Sample size for grade K–3 teachers in all districts = 50 schools and 632 teachers for the treatment group; 50 schools and 668 teachers for the control group. The analyses were based on a two-level analysis (teachers within schools) controlling for random assignment blocks.

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

SOURCE: Spring 2013 Teacher Survey.

Exhibit I.3b. Percentage of teachers who reported receiving specific types of student achievement information, overall and within CLASS and FFT districts, by treatment status, Year 2

	Treatment group mean	Control group mean	Estimated difference	p value
Grade 4–8 teachers in all districts				
Value-added scores for me based upon the students that I taught	80.9	33.5	47.4*	0.000
Data on individual students that I taught	72.8	87.8	-15.0*	0.000
Average data for classes of students that I taught	72.4	72.8	-0.4	0.902
I did not receive any student achievement information based on standardized test results.	8.2	6.3	1.8	0.290
Grade 4–8 teachers in CLASS districts				
Value-added scores for me based upon the students that I taught	76.0	46.9	29.1*	0.000
Data on individual students that I taught	77.8	85.8	-8.0*	0.013
Average data for classes of students that I taught	69.2	71.3	-2.1	0.575
I did not receive any student achievement information based on standardized test results.	7.0	6.4	0.6	0.775
Grade 4–8 teachers in FFT districts				
Value-added scores for me based upon the students that I taught	85.6	21.6	64.0*	0.000
Data on individual students that I taught	68.1	89.7	-21.7*	0.000
Average data for classes of students that I taught	75.5	74.1	1.4	0.777
I did not receive any student achievement information based on standardized test results.	9.3	6.4	2.9	0.330
Grade K–3 teachers in all districts				
Value-added scores for me based upon the students that I taught	42.3	37.1	5.1	0.113
Data on individual students that I taught	67.7	80.7	-13.0*	0.000
Average data for classes of students that I taught	51.7	62.2	-10.4*	0.004
I did not receive any student achievement information based on standardized test results.	25.1	12.0	13.1*	0.000

NOTES: Sample size for grade 4–8 teachers in all districts = 63 schools and 492–498 teachers for the treatment group; 63 schools and 521 or 522 teachers for the control group. Sample size for grade 4–8 teachers in CLASS districts = 31 schools and 296–300 teachers for the treatment group; 32 schools and 311 or 312 teachers for the control group. Sample size for grade 4–8 teachers in FFT districts = 32 schools and 196–198 teachers for the treatment group; 31 schools and 210 teachers for the control group. Sample size for grade K–3 teachers in all districts = 49 schools and 655–662 teachers for the treatment group; 47 schools and 653–655 teachers for the control group.

The analyses were based on a two-level analysis (teachers within schools) controlling for random assignment blocks.

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

SOURCE: Spring 2014 Teacher Survey.

Exhibit I.4a. Number of feedback instances and duration of feedback that an average principal reported receiving, overall and within CLASS and FFT districts, by treatment status, Year 1

	Treatment group median	Control group median	Estimated difference	p value
All districts				
Number of instances of feedback	2.0	1.4	0.6*	0.018
Number of instances of feedback accompanied by a rating	1.0	0.4	0.6*	0.000
Total length of feedback sessions across the year	60.0	40.8	19.2*	0.044
CLASS districts				
Number of instances of feedback	1.0	0.5	0.5	0.180
Number of feedback sessions with ratings	0.0	-0.3	0.3*	0.000
Total length of feedback sessions across the year	25.0	17.5	7.5	0.466
FFT districts				
Number of instances of feedback	2.0	1.3	0.7	0.086
Number of feedback sessions with ratings	1.0	0.2	0.8*	0.000
Total length of feedback sessions across the year	90.0	59.9	30.1*	0.046

NOTES: Sample size for all districts = 61 principals for the treatment group; 61 principals for the control group. Sample size for CLASS districts = 31 principals for the treatment group; 30 principals for the control group. Sample size for FFT districts = 30 principals for the treatment group; 31 principals for the control group.

The analyses were based on an aligned rank sum test with randomization inference about median difference between treatment and control groups.

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

SOURCE: Spring 2013 Principal Survey.

Exhibit I.4b. Number of feedback instances and duration of feedback that an average principal reported receiving, overall and within CLASS and FFT districts, by treatment status, Year 2

	Treatment group median	Control group median	Estimated difference	p value
All districts				
Number of instances of feedback	3.0	2.5	0.5	0.166
Number of instances of feedback accompanied by a rating	2.0	1.0	1.0*	0.000
Total length of feedback sessions across the year	60.0	32.9	27.1*	0.022
CLASS districts				
Number of instances of feedback	2.0	2.0	0.0	0.922
Number of feedback sessions with ratings	0.0	-0.4	0.4	0.072
Total length of feedback sessions across the year	30.0	31.2	-1.2	0.896
FFT districts				
Number of instances of feedback	3.0	2.0	1.0*	0.042
Number of feedback sessions with ratings	2.0	0.5	1.5*	0.000
Total length of feedback sessions across the year	105.0	45.0	60.0*	0.000

NOTES: Sample size for all districts = 61 principals for the treatment group; 59 principals for the control group. Sample size for CLASS districts = 29 principals for the treatment group; 29 principals for the control group. Sample size for FFT districts = 32 principals for the treatment group; 30 principals for the control group.

The analyses were based on an aligned rank sum test with randomization inference about median difference between treatment and control groups.

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

SOURCE: Spring 2014 Principal Survey.

Supporting Exhibits for Analyses of Initial Outcomes

Exhibit I.5a. Percentage of teachers who reported discussing areas of practice related to CLASS/FFT with someone who provided them with feedback during the school year, overall and within CLASS and FFT districts, by treatment status, Year 1

Area of practice	Treatment group mean	Control group mean	Estimated difference	p value
Grade 4–8 teachers in all districts				
Discussed at least one CLASS/FFT area	86.7	72.7	14.0*	0.000
Behavior management	56.4	51.3	5.2	0.174
Classroom organization	52.4	39.5	13.0*	0.000
Emotional support for students	50.4	39.4	11.0*	0.001
Instructional dialogue	72.0	54.3	17.7*	0.000
Student engagement	73.6	52.9	20.7*	0.000
Grade 4–8 teachers in CLASS districts				
Discussed at least one CLASS/FFT area	82.3	74.4	7.9*	0.036
Behavior management	50.7	51.8	-1.1	0.819
Classroom organization	45.9	39.9	6.0	0.227
Emotional support for students	54.6	40.1	14.5*	0.001
Instructional dialogue	70.7	50.5	20.2*	0.000
Student engagement	66.1	50.0	16.1*	0.001
Grade 4–8 teachers in FFT districts				
Discussed at least one CLASS/FFT area	90.9	71.2	19.7*	0.000
Behavior management	62.0	51.1	10.9	0.082
Classroom organization	58.8	39.2	19.6*	0.000
Emotional support for students	46.4	38.4	8.0	0.168
Instructional dialogue	73.2	58.1	15.1*	0.005
Student engagement	80.9	55.7	25.1*	0.000
Grade K–3 teachers in all districts				
Discussed at least one CLASS/FFT area	87.3	75.7	11.6*	0.000
Behavior management	60.1	60.6	-0.5	0.891
Classroom organization	53.8	44.5	9.3*	0.017
Emotional support for students	50.0	41.9	8.1*	0.047
Instructional dialogue	72.0	53.4	18.6*	0.000
Student engagement	69.5	55.6	13.9*	0.000

NOTES: Sample size for grade 4–8 teachers in all districts = 63 schools and 453–463 teachers for the treatment group; 64 schools and 477–488 teachers for the control group. Sample size for grade 4–8 teachers in CLASS districts = 31 schools and 264–270 teachers for the treatment group; 32 schools and 271–276 teachers for the control group. Sample size for grade 4–8 teachers in FFT districts = 32 schools and 189–194 teachers for the treatment group; 32 schools and 209–212 teachers for the control group. Sample size for grade K–3 teachers in all districts = 50 schools and 564–577 teachers for the treatment group; 50 schools and 576–595 teachers for the control group. The analyses were based on a two-level analysis (teachers within schools) controlling for random assignment blocks and teacher background characteristics.

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

SOURCE: Spring 2013 Teacher Survey.

Exhibit I.5b. Percentage of teachers who reported discussing areas of practice related to CLASS/FFT with someone who provided them with feedback during the school year, overall and within CLASS and FFT districts, by treatment status, Year 2

Area of practice	Treatment group mean	Control group mean	Estimated difference	p value
Grade 4–8 teachers in all districts				
Discussed at least one CLASS/FFT area	89.4	77.9	11.5*	0.000
Behavior management	62.5	53.6	8.9*	0.027
Classroom organization	54.5	42.0	12.5*	0.001
Emotional support for students	54.9	42.4	12.5*	0.002
Instructional dialogue	74.7	55.1	19.6*	0.000
Student engagement	73.2	59.0	14.2*	0.000
Grade 4–8 teachers in CLASS districts				
Discussed at least one CLASS/FFT area	88.6	73.4	15.2*	0.000
Behavior management	59.1	51.5	7.6	0.142
Classroom organization	51.7	39.1	12.5*	0.009
Emotional support for students	60.9	39.5	21.4*	0.000
Instructional dialogue	76.3	50.5	25.7*	0.000
Student engagement	70.5	56.9	13.6*	0.001
Grade 4–8 teachers in FFT districts				
Discussed at least one CLASS/FFT area	90.1	82.3	7.8*	0.035
Behavior management	65.8	55.5	10.3	0.098
Classroom organization	57.2	44.6	12.6*	0.026
Emotional support for students	49.2	45.5	3.6	0.527
Instructional dialogue	73.2	59.6	13.5*	0.015
Student engagement	75.8	59.7	16.1*	0.005
Grade K–3 teachers in all districts				
Discussed at least one CLASS/FFT area	87.6	79.4	8.2*	0.002
Behavior management	61.3	59.0	2.3	0.488
Classroom organization	53.7	48.3	5.4	0.146
Emotional support for students	47.6	47.8	-0.3	0.939
Instructional dialogue	73.4	61.4	12.0*	0.000
Student engagement	67.5	59.5	8.0	0.019

NOTES: Sample size for grade 4–8 teachers in all districts = 63 schools and 493–497 teachers for the treatment group; 63 schools and 514–519 teachers for the control group. Sample size for grade 4–8 teachers in CLASS districts = 31 schools and 298 or 299 teachers for the treatment group; 32 schools and 305–310 teachers for the control group. Sample size for grade 4–8 teachers in FFT districts = 32 schools and 195–198 teachers for the treatment group; 31 schools and 209 teachers for the control group.

Sample size for grade K–3 teachers in all districts = 49 schools and 655–662 teachers for the treatment group; 47 schools and 646–656 teachers for the control group. The analyses were based on a two-level analysis (teachers within schools) controlling for random assignment blocks and teacher background characteristics.

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

SOURCE: Spring 2014 Teacher Survey.

Exhibit I.6a. Percentage of teachers who reported discussing areas of practice not related to CLASS/FFT with someone who provided them with feedback during the school year, overall and within CLASS and FFT districts, by treatment status, Year 1

Area of practice	Treatment group mean	Control group mean	Estimated difference	p value
Grade 4–8 teachers in all districts				
Discussed at least one non-CLASS/non-FFT area	73.4	76.9	-3.5	0.270
Lesson planning	46.6	49.4	-2.8	0.428
Data use	57.6	62.5	-4.8	0.210
Content-specific teaching techniques	51.3	53.1	-1.9	0.631
Content knowledge	47.9	50.8	-2.9	0.449
Grade 4–8 teachers in CLASS districts				
Discussed at least one non-CLASS/non-FFT area	70.4	76.0	-5.6	0.193
Lesson planning	49.7	46.8	2.9	0.508
Data use	53.7	60.8	-7.0	0.160
Content-specific teaching techniques	48.0	51.0	-2.9	0.534
Content knowledge	49.3	48.6	0.7	0.891
Grade 4–8 teachers in FFT districts				
Discussed at least one non-CLASS/non-FFT area	76.2	77.7	-1.5	0.750
Lesson planning	43.7	51.4	-7.7	0.205
Data use	61.4	63.7	-2.3	0.698
Content-specific teaching techniques	54.4	55.0	-0.6	0.924
Content knowledge	46.5	52.8	-6.3	0.296
Grade K–3 teachers in all districts				
Discussed at least one non-CLASS/non-FFT area	79.7	78.1	1.6	0.544
Lesson planning	52.0	52.1	-0.1	0.976
Data use	61.8	64.6	-2.8	0.385
Content-specific teaching techniques	51.8	49.3	2.5	0.494
Content knowledge	49.7	50.3	-0.6	0.857

NOTES: Sample size for grade 4–8 teachers in all districts = 63 schools and 453–462 teachers for the treatment group; 64 schools and 470–487 teachers for the control group. Sample size for grade 4–8 teachers in CLASS districts = 31 schools and 262–270 teachers for the treatment group; 32 schools and 263–276 teachers for the control group. Sample size for grade 4–8 teachers in FFT districts = 32 schools and 191–194 teachers for the treatment group; 32 schools and 207–212 teachers for the control group. Sample size for grade K–3 teachers in all districts = 50 schools and 570–578 teachers for the treatment group; 50 schools and 577–592 teachers for the control group. The analyses were based on a two-level analysis (teachers within schools) controlling for random assignment blocks and teacher background characteristics.

None of the differences between the treatment and the control groups are statistically significant at the .05 level (two-tailed).

SOURCE: Spring 2013 Teacher Survey.

Exhibit I.6b. Percentage of teachers who reported discussing areas of practice not related to CLASS/FFT with someone who provided them with feedback during the school year, overall and within CLASS and FFT districts, by treatment status, Year 2

Area of practice	Treatment group mean	Control group mean	Estimated difference	p value
Grade 4–8 teachers in all districts				
Discussed at least one non-CLASS/non-FFT area	75.9	79.7	-3.8	0.172
Lesson planning	53.3	54.5	-1.2	0.730
Data use	66.3	67.8	-1.6	0.670
Content-specific teaching techniques	50.2	51.8	-1.6	0.684
Content knowledge	51.0	51.4	-0.4	0.916
Grade 4–8 teachers in CLASS districts				
Discussed at least one non-CLASS/non-FFT area	77.5	76.7	0.7	0.829
Lesson planning	55.3	53.7	1.6	0.723
Data use	67.5	65.2	2.4	0.600
Content-specific teaching techniques	49.3	49.1	0.2	0.973
Content knowledge	51.5	49.8	1.6	0.739
Grade 4–8 teachers in FFT districts				
Discussed at least one non-CLASS/non-FFT area	74.4	82.2	-7.7	0.086
Lesson planning	51.4	55.7	-4.3	0.403
Data use	65.0	70.2	-5.2	0.376
Content-specific teaching techniques	51.1	54.4	-3.3	0.593
Content knowledge	50.5	52.7	-2.3	0.713
Grade K–3 teachers in all districts				
Discussed at least one non-CLASS/non-FFT area	77.6	81.2	-3.6	0.200
Lesson planning	53.3	54.5	-1.2	0.726
Data use	64.7	69.6	-4.9	0.131
Content-specific teaching techniques	52.6	55.2	-2.6	0.438
Content knowledge	51.6	56.2	-4.6	0.128

NOTES: Sample size for grade 4–8 teachers in all districts = 63 schools and 496 or 497 teachers for the treatment group; 63 schools and 516–519 teachers for the control group. Sample size for grade 4–8 teachers in CLASS districts = 31 schools and 297–299 teachers for the treatment group; 32 schools and 306–310 teachers for the control group. Sample size for grade 4–8 teachers in FFT districts = 32 schools and 198 teachers for the treatment group; 31 schools and 208 or 209 teachers for the control group.

Sample size for grade K–3 teachers in all districts = 49 schools and 652–661 teachers for the treatment group; 47 schools and 648–654 teachers for the control group. The analyses were based on a two-level analysis (teachers within schools) controlling for random assignment blocks and teacher background characteristics.

None of the differences between the treatment and the control groups are statistically significant at the .05 level (two-tailed).

SOURCE: Spring 2014 Teacher Survey.

Exhibit I.7a. Percentage of teachers who reported wanting to improve in areas of practice related to CLASS/FFT by a moderate or large amount, overall and within CLASS and FFT districts, by treatment status, Year 1

Area of practice	Treatment group mean	Control group mean	Estimated difference	p value
Grade 4–8 teachers in all districts				
Interested in improving in at least one CLASS/FFT area	58.8	59.1	-0.3	0.931
Behavior management	29.5	29.4	0.1	0.968
Classroom organization	25.7	27.2	-1.5	0.627
Emotional support for students	26.6	31.1	-4.5	0.184
Instructional dialogue	46.8	45.6	1.2	0.715
Student engagement	40.5	41.3	-0.7	0.849
Grade 4–8 teachers in CLASS districts				
Interested in improving in at least one CLASS/FFT area	56.8	58.9	-2.1	0.657
Behavior management	27.6	30.9	-3.3	0.468
Classroom organization	29.2	28.0	1.2	0.754
Emotional support for students	25.5	31.6	-6.1	0.094
Instructional dialogue	46.6	46.0	0.6	0.881
Student engagement	39.8	41.0	-1.2	0.798
Grade 4–8 teachers in FFT districts				
Interested in improving in at least one CLASS/FFT area	60.7	59.1	1.7	0.775
Behavior management	31.4	28.6	2.8	0.614
Classroom organization	22.2	26.6	-4.4	0.384
Emotional support for students	27.8	28.8	-1.0	0.872
Instructional dialogue	47.0	44.9	2.1	0.713
Student engagement	41.3	40.6	0.7	0.916
Grade K–3 teachers in all districts				
Interested in improving in at least one CLASS/FFT area	62.6	57.6	5.0	0.089
Behavior management	33.6	32.8	0.8	0.798
Classroom organization	29.4	28.0	1.4	0.634
Emotional support for students	28.4	27.4	0.9	0.737
Instructional dialogue	50.1	40.7	9.3*	0.003
Student engagement	39.9	32.7	7.2*	0.013

NOTES: Sample size for grade 4–8 teachers in all districts = 63 schools and 517–521 teachers for the treatment group; 64 schools and 546–552 teachers for the control group. Sample size for grade 4–8 teachers in CLASS districts = 31 schools and 303–305 teachers for the treatment group; 32 schools and 321–324 teachers for the control group. Sample size for grade 4–8 teachers in FFT districts = 32 schools and 214–216 teachers for the treatment group; 32 schools and 224–228 teachers for the control group. Sample size for grade K–3 teachers in all districts = 50 schools and 627–634 teachers for the treatment group; 50 schools and 662–670 teachers for the control group. The analyses were based on a two-level analysis (teachers within schools) controlling for random assignment blocks and teacher background characteristics.

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

SOURCE: Spring 2013 Teacher Survey.

Exhibit I.7b. Percentage of teachers who reported wanting to improve in areas of practice related to CLASS/FFT by a moderate or large amount, overall and within CLASS and FFT districts, by treatment status, Year 2

Area of practice	Treatment group mean	Control group mean	Estimated difference	p value
Grade 4–8 teachers in all districts				
Interested in improving in at least one CLASS/FFT area	57.1	58.1	-1.0	0.779
Behavior management	27.0	27.9	-0.9	0.761
Classroom organization	25.9	24.7	1.2	0.697
Emotional support for students	25.5	27.4	-1.9	0.555
Instructional dialogue	45.2	43.4	1.9	0.595
Student engagement	36.7	40.1	-3.5	0.286
Grade 4–8 teachers in CLASS districts				
Interested in improving in at least one CLASS/FFT area	56.6	58.2	-1.6	0.708
Behavior management	24.7	28.9	-4.2	0.302
Classroom organization	24.7	25.4	-0.7	0.861
Emotional support for students	25.1	28.5	-3.4	0.391
Instructional dialogue	47.0	43.6	3.4	0.420
Student engagement	36.6	40.7	-4.2	0.308
Grade 4–8 teachers in FFT districts				
Interested in improving in at least one CLASS/FFT area	57.5	57.8	-0.2	0.970
Behavior management	29.1	27.1	2.1	0.642
Classroom organization	27.1	24.3	2.8	0.533
Emotional support for students	25.8	25.5	0.3	0.951
Instructional dialogue	43.5	42.8	0.8	0.903
Student engagement	36.7	39.5	-2.8	0.615
Grade K–3 teachers in all districts				
Interested in improving in at least one CLASS/FFT area	58.9	57.9	1.0	0.747
Behavior management	30.3	27.6	2.7	0.316
Classroom organization	23.2	26.3	-3.1	0.246
Emotional support for students	23.4	27.3	-3.8	0.134
Instructional dialogue	47.1	43.9	3.2	0.343
Student engagement	37.6	33.7	3.9	0.190

NOTES: Sample size for grade 4–8 teachers in all districts = 63 schools and 496–498 teachers for the treatment group; 63 schools and 518–520 teachers for the control group. Sample size for grade 4–8 teachers in CLASS districts = 31 schools and 299 or 300 teachers for the treatment group; 32 schools and 305–310 teachers for the control group. Sample size for grade 4–8 teachers in FFT districts = 32 schools and 195–198 teachers for the treatment group; 31 schools and 210 teachers for the control group. Sample size for grade K–3 teachers in all districts = 49 schools and 657–662 teachers for the treatment group; 47 schools and 650–656 teachers for the control group. The analyses were based on a two-level analysis (teachers within schools) controlling for random assignment blocks and teacher background characteristics.

None of the differences between the treatment and the control groups are statistically significant at the .05 level (two-tailed).

SOURCE: Spring 2014 Teacher Survey.

Exhibit I.8a. Percentage of teachers who reported wanting to improve in areas of practice not related to CLASS/FFT by a moderate or large amount, overall and within CLASS and FFT districts, by treatment status, Year 1

Area of practice	Treatment group mean	Control group mean	Estimated difference	p value
Grade 4–8 teachers in all districts				
Interested in improving in at least one non-CLASS/non-FFT area	57.8	64.8	-7.0*	0.040
Lesson planning	29.3	32.5	-3.2	0.364
Data use	44.2	50.2	-6.0	0.081
Content-specific teaching techniques	37.0	39.4	-2.3	0.496
Content knowledge	30.8	33.8	-2.9	0.392
Grade 4–8 teachers in CLASS districts				
Interested in improving in at least one non-CLASS/non-FFT area	56.2	62.4	-6.2	0.129
Lesson planning	28.5	32.5	-3.9	0.296
Data use	44.1	47.1	-3.1	0.447
Content-specific teaching techniques	38.0	37.9	0.2	0.966
Content knowledge	33.7	34.2	-0.5	0.896
Grade 4–8 teachers in FFT districts				
Interested in improving in at least one non-CLASS/non-FFT area	59.3	66.4	-7.0	0.256
Lesson planning	30.1	32.1	-1.9	0.753
Data use	44.4	52.7	-8.3	0.167
Content-specific teaching techniques	36.0	40.7	-4.7	0.446
Content knowledge	28.1	33.0	-4.9	0.410
Grade K–3 teachers in all districts				
Interested in improving in at least one non-CLASS/non-FFT area	64.6	61.5	3.2	0.337
Lesson planning	32.9	30.4	2.5	0.380
Data use	46.0	44.4	1.6	0.610
Content-specific teaching techniques	44.4	42.6	1.7	0.561
Content knowledge	39.1	38.6	0.6	0.860

NOTES: Sample size for grade 4–8 teachers in all districts = 63 schools and 509–519 teachers for the treatment group; 64 schools and 531–549 teachers for the control group. Sample size for grade 4–8 teachers in CLASS districts = 31 schools and 297–305 teachers for the treatment group; 32 schools and 309–323 teachers for the control group. Sample size for grade 4–8 teachers in FFT districts = 32 schools and 212–215 teachers for the treatment group; 32 schools and 222–227 teachers for the control group. Sample size for grade K–3 teachers in all districts = 50 schools and 622–631 teachers for the treatment group; 50 schools and 649–667 teachers for the control group. The analyses were based on a two-level analysis (teachers within schools) controlling for random assignment blocks and teacher background characteristics.

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

SOURCE: Spring 2013 Teacher Survey.

Exhibit I.8b. Percentage of teachers who reported wanting to improve in areas of practice not related to CLASS/FFT by a moderate or large amount, overall and within CLASS and FFT districts, by treatment status, Year 2

Area of practice	Treatment group mean	Control group mean	Estimated difference	p value
Grade 4–8 teachers in all districts				
Interested in improving in at least one non-CLASS/non-FFT area	55.5	59.9	-4.4	0.249
Lesson planning	30.0	29.2	0.7	0.827
Data use	41.1	45.0	-3.9	0.288
Content-specific teaching techniques	36.2	38.6	-2.4	0.560
Content knowledge	28.2	34.7	-6.5	0.071
Grade 4–8 teachers in CLASS districts				
Interested in improving in at least one non-CLASS/non-FFT area	55.7	60.1	-4.4	0.385
Lesson planning	30.3	30.1	0.2	0.969
Data use	40.3	45.9	-5.6	0.176
Content-specific teaching techniques	39.0	37.6	1.4	0.795
Content knowledge	31.5	35.2	-3.7	0.437
Grade 4–8 teachers in FFT districts				
Interested in improving in at least one non-CLASS/non-FFT area	55.3	59.7	-4.5	0.435
Lesson planning	29.7	28.4	1.3	0.792
Data use	41.9	44.7	-2.8	0.663
Content-specific teaching techniques	33.5	39.5	-6.0	0.336
Content knowledge	25.0	34.5	-9.5	0.067
Grade K–3 teachers in all districts				
Interested in improving in at least one non-CLASS/non-FFT area	57.8	59.0	-1.2	0.704
Lesson planning	28.2	27.4	0.8	0.771
Data use	41.3	44.8	-3.5	0.247
Content-specific teaching techniques	39.0	35.3	3.7	0.198
Content knowledge	36.9	34.6	2.3	0.425

NOTES: Sample size for grade 4–8 teachers in all districts = 63 schools and 496–498 teachers for the treatment group; 63 schools and 516–520 teachers for the control group. Sample size for grade 4–8 teachers in CLASS districts = 31 schools and 298–300 teachers for the treatment group; 32 schools and 307–310 teachers for the control group. Sample size for grade 4–8 teachers in FFT districts = 32 schools and 198 teachers for the treatment group; 31 schools and 209 or 210 teachers for the control group. Sample size for grade K–3 teachers in all districts = 49 schools and 654–661 teachers for the treatment group; 47 schools and 650–655 teachers for the control group. The analyses were based on a two-level analysis (teachers within schools) controlling for random assignment blocks and teacher background characteristics.

None of the differences between the treatment and the control groups are statistically significant at the .05 level (two-tailed).

SOURCE: Spring 2014 Teacher Survey.

Exhibit I.9a. Percentage of teachers who reported that their professional development activities during Year 1 covered areas of practice related to CLASS/FFT to a moderate or large extent, overall and within CLASS and FFT districts, by treatment status

Area of practice	Treatment group mean	Control group mean	Estimated difference	p value
Grade 4–8 teachers in all districts				
PD focused on at least one CLASS/FFT area	67.7	67.2	0.4	0.884
Behavior management	32.5	26.7	5.7	0.086
Classroom organization	26.6	26.2	0.4	0.891
Emotional support for students	30.0	22.4	7.6*	0.017
Instructional dialogue	51.0	56.8	-5.8	0.091
Student engagement	54.6	52.7	1.9	0.549
Grade 4–8 teachers in CLASS districts				
PD focused on at least one CLASS/FFT area	68.4	66.0	2.4	0.514
Behavior management	35.0	29.7	5.3	0.215
Classroom organization	30.4	29.3	1.1	0.818
Emotional support for students	34.8	27.9	6.9	0.155
Instructional dialogue	50.1	56.5	-6.4	0.188
Student engagement	56.3	54.9	1.4	0.704
Grade 4–8 teachers in FFT districts				
PD focused on at least one CLASS/FFT area	67.0	67.3	-0.3	0.953
Behavior management	30.0	23.7	6.3	0.220
Classroom organization	23.0	23.0	0.0	0.995
Emotional support for students	25.4	17.8	7.7	0.065
Instructional dialogue	51.7	57.0	-5.3	0.282
Student engagement	53.0	49.9	3.2	0.570
Grade K–3 teachers in all districts				
PD focused on at least one CLASS/FFT area	74.9	73.1	1.8	0.530
Behavior management	33.9	29.7	4.2	0.204
Classroom organization	31.9	27.5	4.5	0.117
Emotional support for students	34.5	26.1	8.4*	0.004
Instructional dialogue	62.4	58.7	3.8	0.206
Student engagement	60.9	56.4	4.6	0.188

NOTES: Sample size for grade 4–8 teachers in all districts = 63 schools and 508–511 teachers for the treatment group; 64 schools and 541–545 teachers for the control group. Sample size for grade 4–8 teachers in CLASS districts = 31 schools and 294 or 295 teachers for the treatment group; 32 schools and 316–319 teachers for the control group. Sample size for grade 4–8 teachers in FFT districts = 32 schools and 214–216 teachers for the treatment group; 32 schools and 225 or 226 teachers for the control group. Sample size for grade K–3 teachers in all districts = 50 schools and 622–629 teachers for the treatment group; 50 schools and 662–670 teachers for the control group. The analyses were based on a two-level analysis (teachers within schools) controlling for random assignment blocks and teacher background characteristics.

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

SOURCE: Spring 2013 Teacher Survey.

Exhibit I.9b. Percentage of teachers who reported that their professional development activities during the summer between Years 1 and 2 covered areas of practice related to CLASS/FFT to a moderate or large extent, overall and within CLASS and FFT districts, by treatment status

Area of practice	Treatment group mean	Control group mean	Estimated difference	p value
Grade 4–8 teachers in all districts				
PD focused on at least one CLASS/FFT area	60.9	62.9	-2.0	0.574
Behavior management	26.7	21.7	5.0	0.142
Classroom organization	25.9	25.5	0.4	0.905
Emotional support for students	26.4	21.0	5.4	0.124
Instructional dialogue	46.4	50.3	-3.9	0.290
Student engagement	50.7	47.8	2.9	0.467
Grade 4–8 teachers in CLASS districts				
PD focused on at least one CLASS/FFT area	65.0	63.0	1.9	0.652
Behavior management	28.1	24.3	3.8	0.426
Classroom organization	28.9	30.5	-1.6	0.745
Emotional support for students	32.3	23.4	9.0	0.063
Instructional dialogue	50.7	52.3	-1.6	0.736
Student engagement	53.4	52.9	0.5	0.910
Grade 4–8 teachers in FFT districts				
PD focused on at least one CLASS/FFT area	56.9	61.7	-4.8	0.451
Behavior management	25.3	19.4	5.9	0.182
Classroom organization	23.0	20.3	2.7	0.535
Emotional support for students	20.6	18.7	1.9	0.695
Instructional dialogue	42.2	49.0	-6.8	0.213
Student engagement	48.2	41.7	6.5	0.345
Grade K–3 teachers in all districts				
PD focused on at least one CLASS/FFT area	64.8	66.2	-1.3	0.688
Behavior management	28.9	27.3	1.6	0.640
Classroom organization	32.7	26.3	6.4*	0.044
Emotional support for students	27.4	24.9	2.5	0.440
Instructional dialogue	54.0	50.3	3.7	0.245
Student engagement	51.6	49.4	2.3	0.496

NOTES: Sample size for grade 4–8 teachers in all districts = 63 schools and 439–447 teachers for the treatment group; 63 schools and 433–447 teachers for the control group. Sample size for grade 4–8 teachers in CLASS districts = 31 schools and 268–272 teachers for the treatment group; 32 schools and 259–268 teachers for the control group. Sample size for grade 4–8 teachers in FFT districts = 32 schools and 171–175 teachers for the treatment group; 31 schools and 174–179 teachers for the control group. Sample size for grade K–3 teachers in all districts = 49 schools and 595–616 teachers for the treatment group; 47 schools and 566–579 teachers for the control group.

The analyses were based on a two-level analysis (teachers within schools) controlling for random assignment blocks and teacher background characteristics.

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

SOURCE: Spring 2014 Teacher Survey.

Exhibit I.9c. Percentage of teachers who reported that their professional development activities during Year 2 covered areas of practice related to CLASS/FFT to a moderate or large extent, overall and within CLASS and FFT districts, by treatment status

Area of practice	Treatment group mean	Control group mean	Estimated difference	p value
Grade 4–8 teachers in all districts				
PD focused on at least one CLASS/FFT area	63.0	67.7	-4.8	0.122
Behavior management	28.7	26.1	2.5	0.414
Classroom organization	28.0	26.9	1.1	0.722
Emotional support for students	29.4	22.9	6.5*	0.037
Instructional dialogue	52.2	55.1	-2.9	0.419
Student engagement	53.2	53.9	-0.7	0.841
Grade 4–8 teachers in CLASS districts				
PD focused on at least one CLASS/FFT area	69.4	68.9	0.5	0.896
Behavior management	32.4	29.2	3.3	0.432
Classroom organization	31.6	30.3	1.4	0.741
Emotional support for students	36.0	27.2	8.9*	0.038
Instructional dialogue	57.4	56.7	0.7	0.857
Student engagement	57.8	57.2	0.6	0.882
Grade 4–8 teachers in FFT districts				
PD focused on at least one CLASS/FFT area	56.7	65.7	-9.0	0.078
Behavior management	25.0	23.4	1.6	0.722
Classroom organization	24.5	23.6	0.8	0.848
Emotional support for students	22.9	19.0	3.9	0.360
Instructional dialogue	47.2	51.8	-4.7	0.448
Student engagement	48.7	51.0	-2.2	0.668
Grade K–3 teachers in all districts				
PD focused on at least one CLASS/FFT area	70.3	69.2	1.1	0.710
Behavior management	32.4	30.3	2.1	0.526
Classroom organization	35.3	30.0	5.4	0.062
Emotional support for students	30.9	29.0	1.9	0.551
Instructional dialogue	59.4	57.8	1.6	0.628
Student engagement	56.9	54.4	2.5	0.454

NOTES: Sample size for grade 4–8 teachers in all districts = 63 schools and 479–485 teachers for the treatment group; 63 schools and 506–511 teachers for the control group. Sample size for grade 4–8 teachers in CLASS districts = 31 schools and 280–291 teachers for the treatment group; 32 schools and 297–305 teachers for the control group. Sample size for grade 4–8 teachers in FFT districts = 32 schools and 193 or 194 teachers for the treatment group; 31 schools and 196–207 teachers for the control group. Sample size for grade K–3 teachers in all districts = 49 schools and 628–655 teachers for the treatment group; 47 schools and 628–648 teachers for the control group. The analyses were based on a two-level analysis (teachers within schools) controlling for random assignment blocks and teacher background characteristics.

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

SOURCE: Spring 2014 Teacher Survey.

Exhibit I.10a. Percentage of teachers who reported that their professional development activities during Year 1 covered areas of practice not related to CLASS/FFT to a moderate or large extent, overall and within CLASS and FFT districts, by treatment status

Area of practice	Treatment group mean	Control group mean	Estimated difference	p value
Grade 4–8 teachers in all districts				
PD focused on at least one non-CLASS/non-FFT area	83.2	86.5	-3.3	0.224
Lesson planning	45.6	49.5	-3.9	0.309
Data use	62.3	65.7	-3.4	0.376
Content-specific teaching techniques	49.7	48.9	0.8	0.815
Content knowledge	47.5	50.1	-2.6	0.433
Grade 4–8 teachers in CLASS districts				
PD focused on at least one non-CLASS/non-FFT area	82.6	86.3	-3.7	0.317
Lesson planning	52.3	53.9	-1.6	0.758
Data use	68.0	67.3	0.7	0.903
Content-specific teaching techniques	50.1	47.5	2.5	0.571
Content knowledge	48.7	52.7	-4.0	0.345
Grade 4–8 teachers in FFT districts				
PD focused on at least one non-CLASS/non-FFT area	83.8	87.4	-3.6	0.362
Lesson planning	39.1	45.0	-5.9	0.263
Data use	56.7	64.2	-7.5	0.182
Content-specific teaching techniques	49.3	50.5	-1.2	0.813
Content knowledge	46.3	47.5	-1.2	0.829
Grade K–3 teachers in all districts				
PD focused on at least one non-CLASS/non-FFT area	83.5	86.2	-2.7	0.271
Lesson planning	52.8	50.3	2.4	0.484
Data use	64.4	64.6	-0.2	0.951
Content-specific teaching techniques	54.1	53.3	0.7	0.806
Content knowledge	56.3	55.6	0.7	0.828

NOTES: Sample size for grade 4–8 teachers in all districts = 63 schools and 501–511 teachers for the treatment group; 64 schools and 539–546 teachers for the control group. Sample size for grade 4–8 teachers in CLASS districts = 31 schools and 289–295 teachers for the treatment group; 32 schools and 314–319 teachers for the control group. Sample size for grade 4–8 teachers in FFT districts = 32 schools and 212–216 teachers for the treatment group; 32 schools and 225–227 teachers for the control group. Sample size for grade K–3 teachers in all districts = 50 schools and 616–626 teachers for the treatment group; 50 schools and 654–668 teachers for the control group.

The analyses were based on a two-level analysis (teachers within schools) controlling for random assignment blocks and teacher background characteristics.

None of the differences between the treatment and the control groups are statistically significant at the .05 level (two-tailed).

SOURCE: Spring 2013 Teacher Survey.

Exhibit I.10b. Percentage of teachers who reported that their professional development activities during the summer between Years 1 and 2 covered areas of practice not related to CLASS/FFT to a moderate or large extent, overall and within CLASS and FFT districts, by treatment status

Area of practice	Treatment group mean	Control group mean	Estimated difference	p value
Grade 4–8 teachers in all districts				
PD focused on at least one non-CLASS/non-FFT area	77.4	78.4	-1.1	0.720
Lesson planning	47.8	46.7	1.1	0.768
Data use	50.9	51.4	-0.6	0.869
Content-specific teaching techniques	43.7	44.4	-0.7	0.841
Content knowledge	43.4	43.3	0.2	0.962
Grade 4–8 teachers in CLASS districts				
PD focused on at least one non-CLASS/non-FFT area	82.4	80.8	1.6	0.637
Lesson planning	50.1	52.2	-2.1	0.633
Data use	58.4	56.7	1.8	0.671
Content-specific teaching techniques	48.1	49.2	-1.1	0.808
Content knowledge	44.3	48.6	-4.3	0.355
Grade 4–8 teachers in FFT districts				
PD focused on at least one non-CLASS/non-FFT area	72.5	75.3	-2.9	0.607
Lesson planning	45.6	39.6	6.0	0.377
Data use	43.6	45.8	-2.2	0.683
Content-specific teaching techniques	39.5	40.0	-0.6	0.922
Content knowledge	42.6	38.1	4.5	0.492
Grade K–3 teachers in all districts				
PD focused on at least one non-CLASS/non-FFT area	77.0	78.4	-1.3	0.621
Lesson planning	48.5	45.1	3.5	0.299
Data use	52.2	51.5	0.7	0.844
Content-specific teaching techniques	48.3	46.5	1.8	0.579
Content knowledge	47.7	47.6	0.1	0.971

NOTES: Sample size for grade 4–8 teachers in all districts = 63 schools and 436-446 teachers for the treatment group; 63 schools and 436-448 teachers for the control group. Sample size for grade 4–8 teachers in CLASS districts = 31 schools and 264-272 teachers for the treatment group; 32 schools and 262-267 teachers for the control group. Sample size for grade 4–8 teachers in FFT districts = 32 schools and 172-175 teachers for the treatment group; 31 schools and 174-181 teachers for the control group. Sample size for grade K–3 teachers in all districts = 49 schools and 599-614 teachers for the treatment group; 47 schools and 564-582 teachers for the control group.

The analyses were based on a two-level analysis (teachers within schools) controlling for random assignment blocks and teacher background characteristics.

None of the differences between the treatment and the control groups are statistically significant at the .05 level (two-tailed).

SOURCE: Spring 2014 Teacher Survey.

Exhibit I.10c. Percentage of teachers who reported that their professional development activities during Year 2 covered areas of practice not related to CLASS/FFT to a moderate or large extent, overall and within CLASS and FFT districts, by treatment status

Area of practice	Treatment group mean	Control group mean	Estimated difference	p value
Grade 4–8 teachers in all districts				
PD focused on at least one non-CLASS/non-FFT area	78.6	84.5	-5.9*	0.029
Lesson planning	52.1	46.2	5.9	0.068
Data use	62.3	61.4	0.9	0.822
Content-specific teaching techniques	50.6	50.4	0.2	0.956
Content knowledge	50.9	49.5	1.4	0.685
Grade 4–8 teachers in CLASS districts				
PD focused on at least one non-CLASS/non-FFT area	85.6	86.3	-0.7	0.810
Lesson planning	57.5	53.0	4.5	0.265
Data use	71.3	66.1	5.2	0.278
Content-specific teaching techniques	52.5	53.8	-1.3	0.748
Content knowledge	54.9	53.2	1.7	0.731
Grade 4–8 teachers in FFT districts				
PD focused on at least one non-CLASS/non-FFT area	71.8	82.9	-11.1*	0.019
Lesson planning	46.9	39.3	7.7	0.131
Data use	53.6	56.6	-3.0	0.653
Content-specific teaching techniques	48.6	46.8	1.8	0.717
Content knowledge	47.1	45.8	1.2	0.808
Grade K–3 teachers in all districts				
PD focused on at least one non-CLASS/non-FFT area	83.8	83.0	0.8	0.752
Lesson planning	52.2	48.4	3.8	0.264
Data use	65.9	64.4	1.5	0.646
Content-specific teaching techniques	55.9	55.7	0.2	0.954
Content knowledge	55.4	54.8	0.6	0.852

NOTES: Sample size for grade 4–8 teachers in all districts = 63 schools and 472–484 teachers for the treatment group; 63 schools and 494–510 teachers for the control group. Sample size for grade 4–8 teachers in CLASS districts = 31 schools and 282–290 teachers for the treatment group; 32 schools and 294–304 teachers for the control group. Sample size for grade 4–8 teachers in FFT districts = 32 schools and 192–194 teachers for the treatment group; 31 schools and 198–206 teachers for the control group. Sample size for grade K–3 teachers in all districts = 49 schools and 649–657 teachers for the treatment group; 47 schools and 627–646 teachers for the control group.

The analyses were based on a two-level analysis (teachers within schools) controlling for random assignment blocks and teacher background characteristics.

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

SOURCE: Spring 2014 Teacher Survey.

Exhibit I.11. Teachers' self-appraisal of their effectiveness in boosting students' reading/ELA and mathematics achievement, overall and within CLASS and FFT districts, by treatment status and year

Subject Area	Treatment group mean	Control group mean	Estimated difference	Standard error	Effect size	p value
<u>Year 1</u>						
All districts						
Reading	73.5	74.2	-0.6	1.2	-0.04	0.626
Mathematics	77.7	75.3	2.4*	1.1	0.14	0.027
CLASS districts						
Reading	78.2	74.8	3.4*	1.6	0.20	0.039
Mathematics	80.3	76.9	3.5*	1.4	0.20	0.012
FFT districts						
Reading	69.1	73.5	-4.5*	1.9	-0.27	0.017
Mathematics	75.1	73.8	1.4	1.7	0.08	0.415
<u>Year 2</u>						
All districts						
Reading	72.0	73.4	-1.4	1.2	-0.09	0.231
Mathematics	76.1	76.1	0.0	1.3	0.00	0.986
CLASS districts						
Reading	75.0	74.4	0.6	1.5	0.04	0.687
Mathematics	78.1	77.5	0.6	1.6	0.04	0.716
FFT districts						
Reading	69.1	72.4	-3.3	1.8	-0.19	0.065
Mathematics	74.2	74.8	-0.5	1.9	-0.03	0.788

NOTES: Year 1 sample size for reading in all districts = 63 schools and 428 teachers for the treatment group; 64 schools and 437 teachers for the control group. Year 1 sample size for mathematics in all districts = 63 schools and 425 teachers for the treatment group; 64 schools and 441 teachers for the control group. Year 1 sample size for reading in CLASS districts = 31 schools and 237 teachers for the treatment group; 32 schools and 252 teachers for the control group. Year 1 sample size for mathematics in CLASS districts = 31 schools and 241 teachers for the treatment group; 32 schools and 257 teachers for the control group. Year 1 sample size for reading in FFT districts = 32 schools and 191 teachers for the treatment group; 32 schools and 185 teachers for the control group. Year 1 sample size for mathematics in FFT districts = 32 schools and 184 teachers for the treatment group; 32 schools and 184 teachers for the control group. Year 2 sample size for reading in all districts = 63 schools and 398 teachers for the treatment group; 63 schools and 414 teachers for the control group. Year 2 sample size for mathematics in all districts = 63 schools and 401 teachers for the treatment group; 63 schools and 396 teachers for the control group. Year 2 sample size for reading in CLASS districts = 31 schools and 232 teachers for the treatment group; 32 schools and 244 teachers for the control group. Year 2 sample size for mathematics in CLASS districts = 31 schools and 236 teachers for the treatment group; 32 schools and 239 teachers for the control group. Year 2 sample size for reading in FFT districts = 32 schools and 166 teachers for the treatment group; 31 schools and 170 teachers for the control group. Year 2 sample size for mathematics in FFT districts = 32 schools and 165 teachers for the treatment group; 31 schools and 157 teachers for the control group.

The analyses were based on a two-level analysis (teachers within schools) controlling for random assignment blocks and teacher background characteristics.

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

SOURCES: Spring 2013 and 2014 Teacher Surveys.

Exhibit I.12a. The association between teachers' self-appraisal of their effectiveness in boosting students' reading/ELA and mathematics achievement and their prior-value-added score, overall and within CLASS and FFT districts, by treatment status and year

Subject Area	Association in treatment group	Association in control group	Estimated difference	Standard error	p value
<u>Year 1</u>					
All districts					
Reading	0.10	0.07	0.03	0.04	0.413
Mathematics	0.13	0.10	0.02	0.04	0.513
CLASS districts					
Reading	0.07	0.06	0.01	0.05	0.821
Mathematics	0.12	0.11	0.02	0.05	0.728
FFT districts					
Reading	0.14	0.09	0.05	0.06	0.389
Mathematics	0.13	0.09	0.04	0.06	0.566
<u>Year 2</u>					
All districts					
Reading	0.10	0.09	0.01	0.04	0.767
Mathematics	0.16	0.15	0.01	0.04	0.809
CLASS districts					
Reading	0.11	0.09	0.02	0.06	0.737
Mathematics	0.14	0.12	0.02	0.06	0.696
FFT districts					
Reading	0.09	0.09	0.00	0.07	0.976
Mathematics	0.19	0.22	-0.03	0.07	0.630

NOTES: Year 1 sample size for reading in all districts = 62 schools and 318 teachers for the treatment group; 64 schools and 317 teachers for the control group. Year 1 sample size for mathematics in all districts = 62 schools and 331 teachers for the treatment group; 64 schools and 329 teachers for the control group. Year 1 sample size for reading in CLASS districts = 31 schools and 177 teachers for the treatment group; 32 schools and 180 teachers for the control group. Year 1 sample size for mathematics in CLASS districts = 31 schools and 191 teachers for the treatment group; 32 schools and 196 teachers for the control group. Year 1 sample size for reading in FFT districts = 31 schools and 141 teachers for the treatment group; 32 schools and 137 teachers for the control group. Year 1 sample size for mathematics in FFT districts = 31 schools and 140 teachers for the treatment group; 32 schools and 133 teachers for the control group. Year 2 sample size for reading in all districts = 62 schools and 304 teachers for the treatment group; 60 schools and 284 teachers for the control group. Year 2 sample size for mathematics in all districts = 63 schools and 321 teachers for the treatment group; 62 schools and 298 teachers for the control group. Year 2 sample size for reading in CLASS districts = 31 schools and 173 teachers for the treatment group; 32 schools and 168 teachers for the control group. Year 2 sample size for mathematics in CLASS districts = 31 schools and 187 teachers for the treatment group; 32 schools and 189 teachers for the control group. Year 2 sample size for reading in FFT districts = 31 schools and 131 teachers for the treatment group; 28 schools and 116 teachers for the control group. Year 2 sample size for mathematics in FFT districts = 32 schools and 134 teachers for the treatment group; 30 schools and 109 teachers for the control group. The analyses were based on a two-level analysis (teachers within schools) controlling for random assignment blocks and teacher background characteristics.

None of the differences between the treatment and the control groups are statistically significant at the .05 level (two-tailed).

SOURCES: Spring 2013 and 2014 Teacher Surveys; AIR Value-Added system.

Exhibit I.12b. Teachers' prior value-added percentile for teachers with self-appraisals in different categories, by subject, Year 1

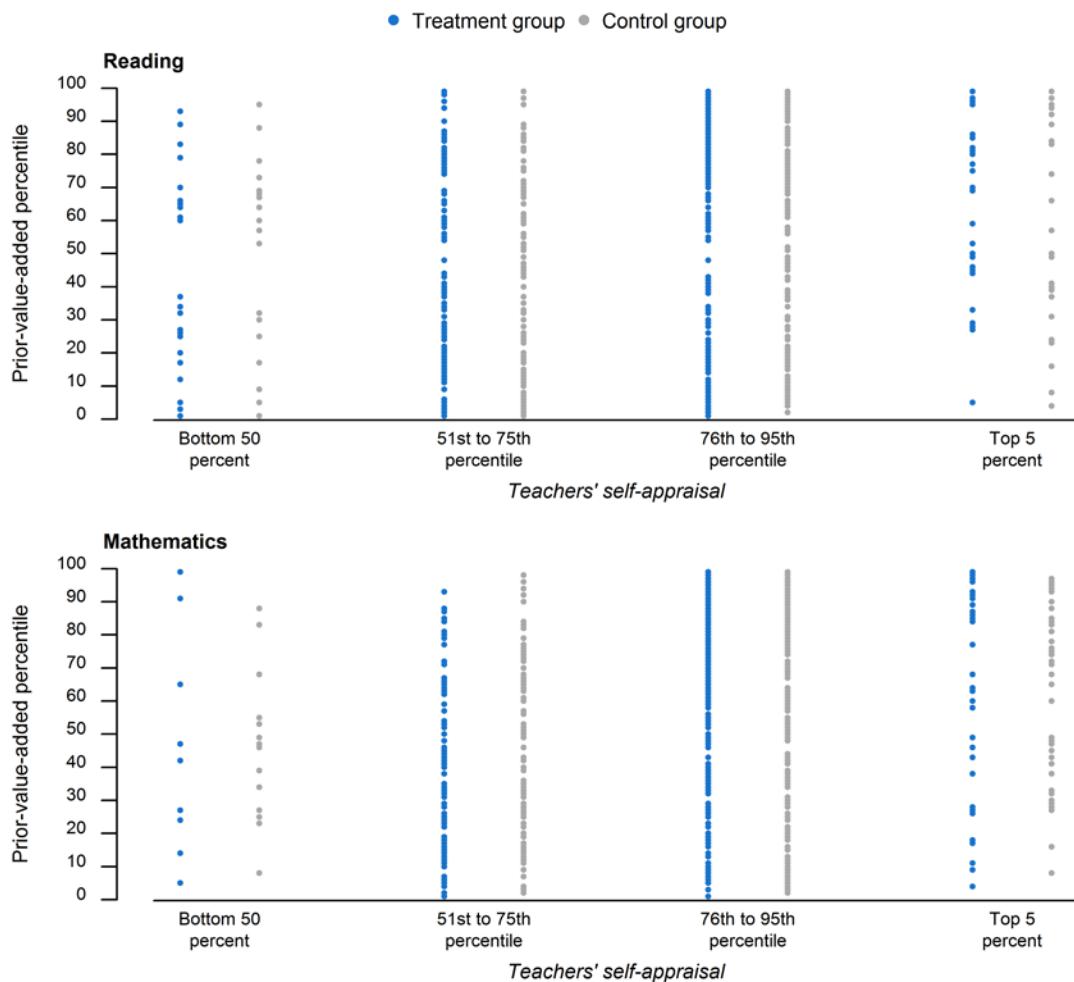


EXHIBIT READS: In Year 1, 14 treatment teachers had a self-rating for reading/ELA in the bottom 50 percent, and a prior-value-added score in the bottom 50 percent among all teachers in the district.

NOTES: Sample size for reading in all districts = 62 schools and 318 teachers for the treatment group; 64 schools and 317 teachers for the control group. Sample size for mathematics in all districts = 62 schools and 331 teachers for the treatment group; 64 schools and 329 teachers for the control group.

SOURCES: Spring 2013 Teacher Surveys; AIR Value-Added system.

Exhibit I.12c. Teachers' prior value-added percentile for teachers with self-appraisals in different categories, by subject, Year 2

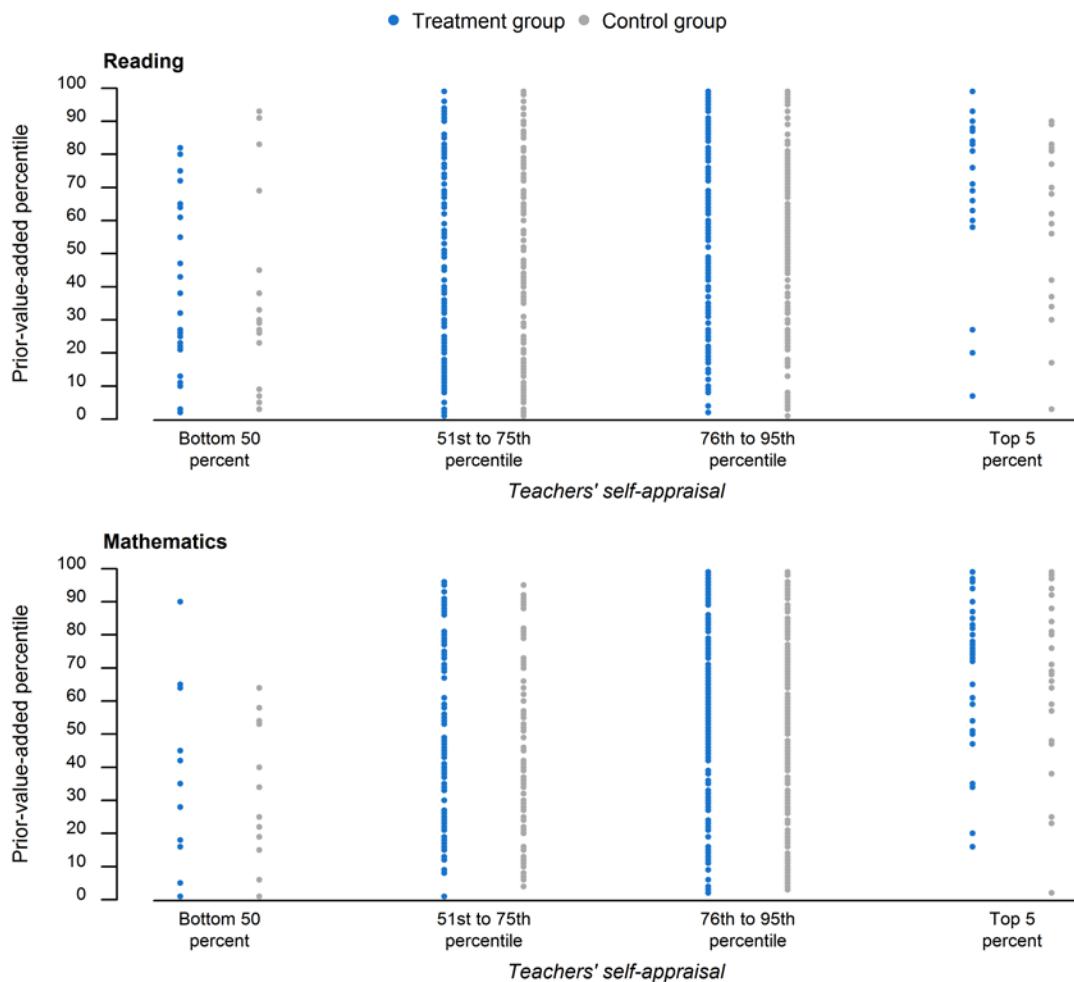


EXHIBIT READS: In Year 2, 17 treatment teachers had a self-rating for reading/ELA in the bottom 50 percent, and a prior-value-added score in the bottom 50 percent among all teachers in the district.

NOTES: Sample size for reading in all districts = 62 schools and 304 teachers for the treatment group; 60 schools and 284 teachers for the control group. Sample size for mathematics in all districts = 63 schools and 321 teachers for the treatment group; 62 schools and 298 teachers for the control group.

SOURCES: Spring 2014 Teacher Surveys; AIR Value-Added system.

Exhibit I.13a. Percentage of principals who reported discussing areas of practice related to the VAL-ED with a supervisor during the school year, overall and within CLASS and FFT districts, by treatment status, Year 1

Area of practice	Treatment group mean	Control group mean	Estimated difference	p value
All districts				
Discussed at least one VAL-ED area	92.1	87.8	4.3	0.427
Identifying, implementing, or monitoring the use of challenging curriculum	52.3	61.7	-9.3	0.305
Advising teachers on ways to improve their instruction	47.8	54.8	-6.9	0.456
Using data to make decisions related to improving student achievement	70.5	64.9	5.6	0.509
Parent/community issues	69.7	46.8	22.9*	0.005
CLASS districts				
Discussed at least one VAL-ED area	†	†	-12.9	0.109
Identifying, implementing, or monitoring the use of challenging curriculum	45.2	69.4	-24.2	0.063
Advising teachers on ways to improve their instruction	45.2	55.0	-9.8	0.421
Using data to make decisions related to improving student achievement	71.0	68.7	2.3	0.852
Parent/community issues	48.4	60.8	-12.4	0.309
FFT districts				
Discussed at least one VAL-ED area	†	†	20.9*	0.006
Identifying, implementing, or monitoring the use of challenging curriculum	59.3	54.2	5.1	0.695
Advising teachers on ways to improve their instruction	50.4	54.6	-4.2	0.768
Using data to make decisions related to improving student achievement	70.0	61.2	8.7	0.449
Parent/community issues	90.3	33.2	57.1*	0.000

NOTES: Sample size for all districts = 63 principals for the treatment group; 64 principals for the control group. Sample size for CLASS districts = 32 principals for the treatment group; 31 principals for the control group. Sample size for FFT districts = 31 principals for the treatment group; 32 principals for the control group.

The analyses were based on a principal-level regression controlling for random assignment blocks and principal background characteristics.

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

†Reporting standards not met; in one or more cells, there are too few cases to report.

SOURCE: Spring 2013 Principal Survey.

Exhibit I.13b. Percentage of principals who reported discussing areas of practice related to the VAL-ED with a supervisor during the school year, overall and within CLASS and FFT districts, by treatment status, Year 2

Area of practice	Treatment group mean	Control group mean	Estimated difference	p value
All districts				
Discussed at least one VAL-ED area	90.0	90.0	0.0	0.999
Identifying, implementing, or monitoring the use of challenging curriculum	62.1	54.2	7.9	0.341
Advising teachers on ways to improve their instruction	61.7	61.6	0.1	0.995
Using data to make decisions related to improving student achievement	70.1	73.4	-3.4	0.707
Parent/community issues	62.8	41.8	21.0*	0.028
CLASS districts				
Discussed at least one VAL-ED area	79.7	91.5	-11.8	0.230
Identifying, implementing, or monitoring the use of challenging curriculum	52.1	64.3	-12.2	0.372
Advising teachers on ways to improve their instruction	47.9	76.3	-28.4	0.053
Using data to make decisions related to improving student achievement	58.5	81.6	-23.1	0.093
Parent/community issues	43.8	37.3	6.4	0.664
FFT districts				
Discussed at least one VAL-ED area	100.0	88.6	11.4	0.081
Identifying, implementing, or monitoring the use of challenging curriculum	71.9	44.5	27.4*	0.009
Advising teachers on ways to improve their instruction	75.0	47.4	27.6*	0.028
Using data to make decisions related to improving student achievement	81.3	65.6	15.7	0.199
Parent/community issues	81.3	46.2	35.1*	0.005

NOTES: Sample size for all districts = 63 principals for the treatment group; 63 principals for the control group. Sample size for CLASS districts = 32 principals for the treatment group; 31 principals for the control group. Sample size for FFT districts = 31 principals for the treatment group; 32 principals for the control group.

The analyses were based on a principal-level regression controlling for random assignment blocks and principal background characteristics.

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

SOURCE: Spring 2014 Principal Survey.

Exhibit I.14a. Percentage of principals who reported discussing areas of practice not related to the VAL-ED with a supervisor during the school year, overall and within CLASS and FFT districts, by treatment status, Year 1

Area of practice	Treatment group mean	Control group mean	Estimated difference	p value
All districts				
Discussed at least one non-VAL-ED area	57.7	71.9	-14.2	0.083
Making personnel/human resources decisions	54.5	53.9	0.6	0.945
Managing nonpersonnel administrative issues	32.7	37.6	-4.9	0.539
Student behavior/discipline	30.9	41.0	-10.1	0.239
CLASS districts				
Discussed at least one non-VAL-ED area	41.9	77.4	-35.5*	0.002
Making personnel/human resources decisions	41.9	61.5	-19.5	0.100
Managing nonpersonnel administrative issues	25.8	45.6	-19.8	0.067
Student behavior/discipline	25.8	41.9	-16.1	0.167
FFT districts				
Discussed at least one non-VAL-ED area	73.0	66.6	6.4	0.597
Making personnel/human resources decisions	66.7	46.6	20.1	0.126
Managing nonpersonnel administrative issues	39.4	29.8	9.6	0.415
Student behavior/discipline	35.8	40.0	-4.2	0.739

NOTES: Sample size for all districts = 63 principals for the treatment group; 64 principals for the control group. Sample size for CLASS districts = 32 principals for the treatment group; 31 principals for the control group. Sample size for FFT districts = 31 principals for the treatment group; 32 principals for the control group.

The analyses were based on a principal-level regression controlling for random assignment blocks and principal background characteristics.

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

SOURCE: Spring 2013 Principal Survey.

Exhibit I.14b. Percentage of principals who reported discussing areas of practice not related to the VAL-ED with a supervisor during the school year, overall and within CLASS and FFT districts, by treatment status, Year 2

Area of practice	Treatment group mean	Control group mean	Estimated difference	p value
All districts				
Discussed at least one non-VAL-ED area	63.7	73.8	-10.1	0.208
Making personnel/human resources decisions	47.8	53.0	-5.1	0.556
Managing nonpersonnel administrative issues	33.8	37.2	-3.4	0.697
Student behavior/discipline	46.9	54.0	-7.1	0.462
CLASS districts				
Discussed at least one non-VAL-ED area	45.6	81.1	-35.4*	0.002
Making personnel/human resources decisions	39.2	66.5	-27.3*	0.030
Managing nonpersonnel administrative issues	23.5	43.5	-19.9	0.092
Student behavior/discipline	37.3	58.0	-20.6	0.140
FFT districts				
Discussed at least one non-VAL-ED area	81.3	66.7	14.5	0.215
Making personnel/human resources decisions	56.3	39.9	16.4	0.190
Managing nonpersonnel administrative issues	43.8	31.1	12.7	0.323
Student behavior/discipline	56.3	50.2	6.1	0.652

NOTES: Sample size for all districts = 63 principals for the treatment group; 63 principals for the control group. Sample size for CLASS districts = 32 principals for the treatment group; 31 principals for the control group. Sample size for FFT districts = 31 principals for the treatment group; 32 principals for the control group.

The analyses were based on a principal-level regression controlling for random assignment blocks and principal background characteristics.

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

SOURCE: Spring 2014 Principal Survey.

Exhibit I.15a. Percentage of principals who reported wanting to improve in areas of practice related to the VAL-ED by a moderate or large amount, overall and within CLASS and FFT districts, by treatment status, Year 1

Area of practice	Treatment group mean	Control group mean	Estimated difference	p value
All districts				
Interested in improving in at least one VAL-ED area	94.9	92.0	2.9	0.518
Identifying, implementing, or monitoring the use of challenging curriculum	65.2	70.0	-4.7	0.571
Advising teachers on ways to improve their instruction	78.5	81.9	-3.3	0.647
Using data to make decisions related to improving student achievement	72.4	77.7	-5.3	0.514
Parent/community issues	67.7	57.5	10.2	0.224
CLASS districts				
Interested in improving in at least one VAL-ED area	†	†	10.1	0.213
Identifying, implementing, or monitoring the use of challenging curriculum	74.2	69.0	5.2	0.681
Advising teachers on ways to improve their instruction	74.2	79.1	-5.0	0.689
Using data to make decisions related to improving student achievement	64.5	74.7	-10.1	0.443
Parent/community issues	58.1	39.9	18.2	0.169
FFT districts				
Interested in improving in at least one VAL-ED area	†	†	-6.8	0.093
Identifying, implementing, or monitoring the use of challenging curriculum	56.5	75.1	-18.6	0.124
Advising teachers on ways to improve their instruction	82.8	89.3	-6.5	0.445
Using data to make decisions related to improving student achievement	80.1	85.8	-5.7	0.577
Parent/community issues	77.1	72.3	4.7	0.685

NOTES: Sample size for all districts = 63 principals for the treatment group; 64 principals for the control group. Sample size for CLASS districts = 32 principals for the treatment group; 31 principals for the control group. Sample size for FFT districts = 31 principals for the treatment group; 32 principals for the control group.

The analyses were based on a principal-level regression controlling for random assignment blocks and principal background characteristics.

None of the differences between the treatment and the control groups are statistically significant at the .05 level (two-tailed).

†Reporting standards not met; in one or more cells, there are too few cases to report.

SOURCE: Spring 2013 Principal Survey.

Exhibit I.15b. Percentage of principals who reported wanting to improve in areas of practice related to the VAL-ED by a moderate or large amount, overall and within CLASS and FFT districts, by treatment status, Year 2

Area of practice	Treatment group mean	Control group mean	Estimated difference	p value
All districts				
Interested in improving in at least one VAL-ED area	90.0	91.9	-1.9	0.725
Identifying, implementing, or monitoring the use of challenging curriculum	67.3	72.7	-5.3	0.511
Advising teachers on ways to improve their instruction	69.6	79.9	-10.3	0.216
Using data to make decisions related to improving student achievement	73.2	77.8	-4.5	0.586
Parent/community issues	60.5	58.2	2.4	0.798
CLASS districts				
Interested in improving in at least one VAL-ED area	†	†	-18.4	0.055
Identifying, implementing, or monitoring the use of challenging curriculum	69.1	72.8	-3.7	0.763
Advising teachers on ways to improve their instruction	57.6	93.5	-35.9*	0.005
Using data to make decisions related to improving student achievement	61.8	89.5	-27.7*	0.039
Parent/community issues	42.4	48.5	-6.1	0.680
FFT districts				
Interested in improving in at least one VAL-ED area	†	†	10.6	0.162
Identifying, implementing, or monitoring the use of challenging curriculum	65.6	71.4	-5.8	0.654
Advising teachers on ways to improve their instruction	81.3	73.5	7.8	0.567
Using data to make decisions related to improving student achievement	84.4	73.9	10.4	0.403
Parent/community issues	78.1	63.6	14.6	0.298

NOTES: Sample size for all districts = 63 principals for the treatment group; 63 principals for the control group. Sample size for CLASS districts = 32 principals for the treatment group; 31 principals for the control group. Sample size for FFT districts = 31 principals for the treatment group; 32 principals for the control group.

The analyses were based on a principal-level regression controlling for random assignment blocks and principal background characteristics.

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

† Reporting standards not met; in one or more cells, there are too few cases to report.

SOURCE: Spring 2014 Principal Survey.

Exhibit I.16a. Percentage of principals who reported wanting to improve in areas of practice not related to the VAL-ED by a moderate or large amount, overall and within CLASS and FFT districts, by treatment status, Year 1

Area of practice	Treatment group mean	Control group mean	Estimated difference	p value
All districts				
Interested in improving in at least one non-VAL-ED area	61.9	59.1	2.8	0.780
Making personnel/human resources decisions	44.7	46.3	-1.6	0.867
Managing nonpersonnel administrative issues	34.2	45.8	-11.6	0.206
Student behavior/discipline	34.9	41.2	-6.3	0.492
CLASS districts				
Interested in improving in at least one non-VAL-ED area	64.5	54.4	10.1	0.487
Making personnel/human resources decisions	54.8	41.1	13.7	0.337
Managing nonpersonnel administrative issues	32.3	50.2	-17.9	0.162
Student behavior/discipline	29.0	34.9	-5.9	0.672
FFT districts				
Interested in improving in at least one non-VAL-ED area	59.3	66.9	-7.6	0.599
Making personnel/human resources decisions	34.9	52.7	-17.7	0.198
Managing nonpersonnel administrative issues	36.1	47.8	-11.6	0.385
Student behavior/discipline	40.5	50.5	-9.9	0.447

NOTES: Sample size for all districts = 63 principals for the treatment group; 64 principals for the control group. Sample size for CLASS districts = 32 principals for the treatment group; 31 principals for the control group. Sample size for FFT districts = 31 principals for the treatment group; 32 principals for the control group.

The analyses were based on a principal-level regression controlling for random assignment blocks and principal background characteristics.

None of the differences between the treatment and the control groups are statistically significant at the .05 level (two-tailed).

SOURCE: Spring 2013 Principal Survey.

Exhibit I.16b. Percentage of principals who reported wanting to improve in areas not related to the VAL-ED by a moderate or large amount, overall and within CLASS and FFT districts, by treatment status, Year 2

Area of practice	Treatment group mean	Control group mean	Estimated difference	p value
All districts				
Interested in improving in at least one non-VAL-ED area	65.3	75.9	-10.6	0.261
Making personnel/human resources decisions	47.4	53.1	-5.7	0.578
Managing nonpersonnel administrative issues	33.3	34.4	-1.0	0.910
Student behavior/discipline	45.4	55.8	-10.4	0.288
CLASS districts				
Interested in improving in at least one non-VAL-ED area	58.5	70.7	-12.2	0.424
Making personnel/human resources decisions	47.9	56.3	-8.3	0.606
Managing nonpersonnel administrative issues	22.6	36.0	-13.4	0.335
Student behavior/discipline	40.6	54.9	-14.4	0.367
FFT districts				
Interested in improving in at least one non-VAL-ED area	71.9	87.1	-15.2	0.283
Making personnel/human resources decisions	46.9	59.9	-13.0	0.392
Managing nonpersonnel administrative issues	43.8	43.2	0.5	0.972
Student behavior/discipline	50.0	58.6	-8.6	0.554

NOTES: Sample size for all districts = 63 principals for the treatment group; 63 principals for the control group. Sample size for CLASS districts = 32 principals for the treatment group; 31 principals for the control group. Sample size for FFT districts = 31 principals for the treatment group; 32 principals for the control group.

The analyses were based on a principal-level regression controlling for random assignment blocks and principal background characteristics.

None of the differences between the treatment and the control groups are statistically significant at the .05 level (two-tailed).

SOURCE: Spring 2014 Principal Survey.

Exhibit I.17a. Percentage of principals who reported that their professional development activities during Year 1 covered areas related to the VAL-ED to a moderate or large extent, overall and within CLASS and FFT districts, by treatment status

Area of practice	Treatment group mean	Control group mean	Estimated difference	p value
All districts				
PD focused on at least one VAL-ED area	†	†	-6.7	0.119
Identifying, implementing, or monitoring the use of challenging curriculum	66.6	72.5	-5.9	0.457
Advising teachers on ways to improve their instruction	68.4	77.1	-8.7	0.263
Using data to make decisions related to improving student achievement	79.9	90.3	-10.4	0.081
Parent/community issues	31.4	28.7	2.6	0.765
CLASS districts				
PD focused on at least one VAL-ED area	100.0	100.0	0.0	1.000
Identifying, implementing, or monitoring the use of challenging curriculum	77.4	75.7	1.8	0.876
Advising teachers on ways to improve their instruction	71.0	86.8	-15.8	0.132
Using data to make decisions related to improving student achievement	†	†	10.3	0.131
Parent/community issues	35.5	31.4	4.1	0.777
FFT districts				
PD focused on at least one VAL-ED area	†	†	-14.4	0.104
Identifying, implementing, or monitoring the use of challenging curriculum	56.0	73.6	-17.6	0.147
Advising teachers on ways to improve their instruction	66.0	68.5	-2.5	0.836
Using data to make decisions related to improving student achievement	†	†	-26.4*	0.011
Parent/community issues	27.4	21.2	6.2	0.577

NOTES: Sample size for all districts = 63 principals for the treatment group; 64 principals for the control group. Sample size for CLASS districts = 32 principals for the treatment group; 31 principals for the control group. Sample size for FFT districts = 31 principals for the treatment group; 32 principals for the control group.

The analyses were based on a principal-level regression controlling for random assignment blocks and principal background characteristics.

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

† Reporting standards not met; in one or more cells, there are too few cases to report.

SOURCE: Spring 2013 Principal Survey.

Exhibit I.17b. Percentage of principals who reported that their professional development activities during the summer between Years 1 and 2 covered areas related to the VAL-ED to a moderate or large extent, overall and within CLASS and FFT districts, by treatment status

Area of practice	Treatment group mean	Control group mean	Estimated difference	p value
All districts				
PD focused on at least one VAL-ED area	88.4	88.8	-0.3	.958
Identifying, implementing, or monitoring the use of challenging curriculum	49.4	55.2	-5.7	.531
Advising teachers on ways to improve their instruction	70.1	70.3	-0.2	.982
Using data to make decisions related to improving student achievement	80.3	78.7	1.6	.839
Parent/community issues	23.8	19.9	3.9	.619
CLASS districts				
PD focused on at least one VAL-ED area	†	†	-8.0	.181
Identifying, implementing, or monitoring the use of challenging curriculum	52.1	70.6	-18.5	.134
Advising teachers on ways to improve their instruction	71.4	74.0	-2.5	.857
Using data to make decisions related to improving student achievement	82.9	88.3	-5.4	.563
Parent/community issues	16.1	16.5	-0.3	.975
FFT districts				
PD focused on at least one VAL-ED area	†	†	12.0	.320
Identifying, implementing, or monitoring the use of challenging curriculum	46.9	46.3	0.5	.972
Advising teachers on ways to improve their instruction	68.8	72.7	-4.0	.782
Using data to make decisions related to improving student achievement	77.7	66.1	11.6	.415
Parent/community issues	31.3	22.2	9.1	.502

NOTES: Sample size for all districts = 63 principals for the treatment group; 63 principals for the control group. Sample size for CLASS districts = 32 principals for the treatment group; 31 principals for the control group. Sample size for FFT districts = 31 principals for the treatment group; 32 principals for the control group.

The analyses were based on a principal-level regression controlling for random assignment blocks and principal background characteristics.

None of the differences between the treatment and the control groups are statistically significant at the .05 level (two-tailed).

† Reporting standards not met; in one or more cells, there are too few cases to report.

SOURCE: Spring 2014 Principal Survey.

Exhibit I.17c. Percentage of principals who reported that their professional development activities during Year 2 covered areas related to the VAL-ED to a moderate or large extent, overall and within CLASS and FFT districts, by treatment status

Area of practice	Treatment group mean	Control group mean	Estimated difference	p value
All districts				
PD focused on at least one VAL-ED area	92.1	90.9	1.2	.825
Identifying, implementing, or monitoring the use of challenging curriculum	64.7	64.0	0.6	.944
Advising teachers on ways to improve their instruction	76.2	87.8	-11.6	.142
Using data to make decisions related to improving student achievement	82.5	83.2	-0.7	.919
Parent/community issues	29.0	25.5	3.5	.720
CLASS districts				
PD focused on at least one VAL-ED area	†	†	-3.2	.461
Identifying, implementing, or monitoring the use of challenging curriculum	66.9	82.3	-15.4	.180
Advising teachers on ways to improve their instruction	†	†	-22.3*	.043
Using data to make decisions related to improving student achievement	†	†	-3.3	.601
Parent/community issues	20.3	20.7	-0.4	.977
FFT districts				
PD focused on at least one VAL-ED area	†	†	6.2	.523
Identifying, implementing, or monitoring the use of challenging curriculum	62.5	47.0	15.5	.324
Advising teachers on ways to improve their instruction	†	†	-2.8	.817
Using data to make decisions related to improving student achievement	†	†	1.6	.899
Parent/community issues	37.5	31.6	5.9	.706

NOTES: Sample size for all districts = 63 principals for the treatment group; 63 principals for the control group. Sample size for CLASS districts = 32 principals for the treatment group; 31 principals for the control group. Sample size for FFT districts = 31 principals for the treatment group; 32 principals for the control group.

The analyses were based on a principal-level regression controlling for random assignment blocks and principal background characteristics.

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

† Reporting standards not met; in one or more cells, there are too few cases to report.

SOURCE: Spring 2014 Principal Survey.

Exhibit I.18a. Percentage of principals who reported that their professional development activities during Year 1 covered areas of practice not related to the VAL-ED to a moderate or large extent, overall and within CLASS and FFT districts, by treatment status

Area of practice	Treatment group mean	Control group mean	Estimated difference	p value
All districts				
PD focused on at least one non-VAL-ED area	50.7	56.4	-5.7	0.532
Making personnel/human resources decisions	30.8	28.3	2.5	0.766
Managing nonpersonnel administrative issues	30.8	30.3	0.4	0.960
Student behavior/discipline	39.4	31.1	8.4	0.309
CLASS districts				
PD focused on at least one non-VAL-ED area	54.8	57.3	-2.5	0.845
Making personnel/human resources decisions	35.5	28.2	7.3	0.572
Managing nonpersonnel administrative issues	38.7	23.7	15.0	0.254
Student behavior/discipline	41.9	34.3	7.6	0.533
FFT districts				
PD focused on at least one non-VAL-ED area	46.7	47.5	-0.8	0.951
Making personnel/human resources decisions	26.3	16.9	9.5	0.342
Managing nonpersonnel administrative issues	23.1	28.2	-5.1	0.660
Student behavior/discipline	37.0	21.9	15.1	0.195

NOTES: Sample size for all districts = 63 principals for the treatment group; 64 principals for the control group. Sample size for CLASS districts = 32 principals for the treatment group; 31 principals for the control group. Sample size for FFT districts = 31 principals for the treatment group; 32 principals for the control group.

The analyses were based on a principal-level regression controlling for random assignment blocks and principal background characteristics.

None of the differences between the treatment and the control groups are statistically significant at the .05 level (two-tailed).

SOURCE: Spring 2013 Principal Survey.

Exhibit I.18b. Percentage of principals who reported that their professional development activities during the summer between Year 1 and 2 covered areas of practice not related to the VAL-ED to a moderate or large extent, overall and within CLASS and FFT districts, by treatment status

Area of practice	Treatment group mean	Control group mean	Estimated difference	p value
All districts				
PD focused on at least one non-VAL-ED area	37.7	38.3	-0.6	0.957
Making personnel/human resources decisions	20.8	25.6	-4.8	0.551
Managing nonpersonnel administrative issues	21.1	22.4	-1.3	0.880
Student behavior/discipline	27.4	29.6	-2.2	0.817
CLASS districts				
PD focused on at least one non-VAL-ED area	28.3	31.9	-3.6	0.804
Making personnel/human resources decisions	13.2	21.5	-8.3	0.385
Managing nonpersonnel administrative issues	20.3	15.4	4.8	0.710
Student behavior/discipline	20.3	23.7	-3.4	0.792
FFT districts				
PD focused on at least one non-VAL-ED area	46.9	46.5	0.4	0.984
Making personnel/human resources decisions	28.1	31.2	-3.1	0.835
Managing nonpersonnel administrative issues	21.9	37.3	-15.4	0.261
Student behavior/discipline	34.4	32.8	1.6	0.922

NOTES: Sample size for all districts = 63 principals for the treatment group; 63 principals for the control group. Sample size for CLASS districts = 32 principals for the treatment group; 31 principals for the control group. Sample size for FFT districts = 31 principals for the treatment group; 32 principals for the control group.

The analyses were based on a principal-level regression controlling for random assignment blocks and principal background characteristics.

None of the differences between the treatment and the control groups are statistically significant at the .05 level (two-tailed).

SOURCE: Spring 2014 Principal Survey.

Exhibit I.18c. Percentage of principals who reported that their professional development activities during Year 2 covered areas not related to the VAL-ED to a moderate or large extent, overall and within CLASS and FFT districts, by treatment status

Area of practice	Treatment group mean	Control group mean	Estimated difference	p value
All districts				
PD focused on at least one non-VAL-ED area	53.1	39.3	13.8	0.209
Making personnel/human resources decisions	33.1	26.8	6.3	0.497
Managing nonpersonnel administrative issues	21.5	26.6	-5.1	0.517
Student behavior/discipline	33.8	32.5	1.3	0.895
CLASS districts				
PD focused on at least one non-VAL-ED area	53.0	31.6	21.4	0.221
Making personnel/human resources decisions	28.6	27.3	1.3	0.931
Managing nonpersonnel administrative issues	18.0	22.0	-4.0	0.723
Student behavior/discipline	30.0	28.0	1.9	0.891
FFT districts				
PD focused on at least one non-VAL-ED area	53.1	55.8	-2.7	0.874
Making personnel/human resources decisions	37.5	36.6	0.9	0.952
Managing nonpersonnel administrative issues	25.0	35.6	-10.6	0.413
Student behavior/discipline	37.5	34.8	2.7	0.863

NOTES: Sample size for all districts = 63 principals for the treatment group; 63 principals for the control group. Sample size for CLASS districts = 32 principals for the treatment group; 31 principals for the control group. Sample size for FFT districts = 31 principals for the treatment group; 32 principals for the control group.

The analyses were based on a principal-level regression controlling for random assignment blocks and principal background characteristics.

None of the differences between the treatment and the control groups are statistically significant at the .05 level (two-tailed).

SOURCE: Spring 2014 Principal Survey.

Exhibit I.19. Principals' self-appraisal of their effectiveness in instructional leadership and other forms of leadership, overall and within CLASS and FFT districts, overall and within CLASS and FFT districts, by treatment status, Year 2

Leadership measure	Treatment group mean	Control group mean	Estimated difference	Standard error	Effect size	p value
All districts						
Instructional leadership	75.7	73.2	2.4	3.0	0.19	0.411
Other forms of leadership	80.4	79.2	1.2	2.3	0.09	0.615
CLASS districts						
Instructional leadership	76.2	77.4	-1.2	4.2	-0.13	0.778
Other forms of leadership	82.0	80.9	1.0	3.3	0.09	0.758
FFT districts						
Instructional leadership	75.2	69.2	6.0	4.3	0.36	0.168
Other forms of leadership	78.9	77.5	1.3	3.4	0.10	0.699

NOTES: Sample size for all districts = 61 principals for the treatment group; 59-60 principals for the control group. Sample size for CLASS districts = 29 principals for the treatment group; 30 principals for the control group. Sample size for FFT districts = 32 principals for the treatment group; 29 or 30 principals for the control group.

The analyses were based on a three-level analysis (observations within teachers within schools) controlling for random assignment blocks and principal background characteristics.

None of the differences between the treatment and the control groups are statistically significant at the .05 level (two-tailed).

SOURCE: Spring 2014 Principal Survey.

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

Appendix J. Supporting Exhibits for Impact Analyses

Supporting Exhibits for Analyses of Baseline Equivalence of the Impact Analysis Samples

Exhibit J.1. Background characteristics of teachers in the Year 2 teacher practice impact sample, overall and within CLASS and FFT districts, by treatment status

Characteristic	Treatment group	Control group	Estimated difference	p value
All districts				
Year of teaching experience				
Years of experience in district	10.5	10.3	0.3	0.709
Mean number of years	13.1	13.0	0.1	0.859
Three years or fewer (percentage)	13.7	18.0	-4.4	0.123
Four to 10 years (percentage)	35.9	30.0	5.9	0.070
Eleven to 20 years (percentage)	28.6	28.8	-0.2	0.943
More than 20 years (percentage)	21.8	23.1	-1.3	0.674
Master's degree or higher (percentage)	45.9	44.5	1.5	0.665
CLASS districts				
Year of teaching experience				
Years of experience in district	11.0	9.9	1.1	0.281
Mean number of years	13.4	12.3	1.1	0.339
Three years or fewer (percentage)	14.2	19.4	-5.2	0.199
Four to 10 years (percentage)	35.7	30.3	5.4	0.212
Eleven to 20 years (percentage)	27.1	30.8	-3.7	0.396
More than 20 years (percentage)	23.1	19.6	3.5	0.384
Master's degree or higher (percentage)	34.3	34.3	0.1	0.985
FFT districts				
Year of teaching experience				
Years of experience in district	10.1	10.7	-0.6	0.570
Mean number of years	12.8	13.8	-0.9	0.358
Three years or fewer (percentage)	13.2	16.5	-3.3	0.377
Four to 10 years (percentage)	36.2	29.8	6.4	0.197
Eleven to 20 years (percentage)	30.0	27.0	3.0	0.567
More than 20 years (percentage)	20.6	26.5	-5.9	0.217
Master's degree or higher (percentage)	57.2	53.6	3.6	0.560

NOTES: Sample size for all districts = 61 schools and 431 teachers for the treatment group; 63 schools and 509 teachers for the control group. Sample size for CLASS districts = 30 schools and 236 teachers for the treatment group; 32 schools and 301 teachers for the control group. Sample size for FFT districts = 31 schools and 195 teachers for the treatment group; 31 schools and 208 teachers for the control group.

The analyses were based on a two-level linear regression (teachers within schools) controlling for random assignment blocks. None of the differences between the treatment and the control groups are statistically significant at the .05 level (two-tailed).

SOURCES: Spring 2014 Teacher Survey.

Exhibit J.2. Background characteristics of teachers in the Year 1 principal leadership impact sample, overall and within CLASS and FFT districts, by treatment status

Characteristic	Treatment group	Control group	Estimated difference	p value
All districts				
Year of teaching experience				
Years of experience in district	10.8	11.2	-0.3	0.596
Mean number of years	13.6	14.0	-0.4	0.559
Three years or fewer (percentage)	12.9	16.3	-3.4	0.176
Four to 10 years (percentage)	35.8	29.2	6.6*	0.045
Eleven to 20 years (percentage)	26.5	28.4	-1.9	0.515
More than 20 years (percentage)	24.7	25.7	-0.9	0.754
Master's degree or higher (percentage)	46.1	45.2	0.8	0.756
CLASS districts				
Year of teaching experience				
Years of experience in district	11.9	10.6	1.2	0.159
Mean number of years	14.1	12.9	1.2	0.191
Three years or fewer (percentage)	10.9	17.9	-7.0*	0.024
Four to 10 years (percentage)	34.0	30.3	3.7	0.432
Eleven to 20 years (percentage)	27.6	29.3	-1.7	0.650
More than 20 years (percentage)	27.5	21.0	6.5	0.093
Master's degree or higher (percentage)	35.2	35.6	-0.5	0.876
FFT districts				
Year of teaching experience				
Years of experience in district	9.8	11.8	-1.9*	0.034
Mean number of years	13.0	15.1	-2.1*	0.036
Three years or fewer (percentage)	14.8	14.9	-0.1	0.982
Four to 10 years (percentage)	37.6	27.7	10.0*	0.033
Eleven to 20 years (percentage)	25.5	27.5	-2.0	0.671
More than 20 years (percentage)	22.1	30.2	-8.1	0.077
Master's degree or higher (percentage)	56.6	54.1	2.5	0.620

NOTES: Sample size for all districts = 63 schools and 524 teachers for the treatment group; 64 schools and 558 teachers for the control group. Sample size for CLASS districts = 31 schools and 306 teachers for the treatment group; 32 schools and 329 teachers for the control group. Sample size for FFT districts = 32 schools and 218 teachers for the treatment group; 32 schools and 229 teachers for the control group.

The analyses were based on a two-level regression (teachers within schools) controlling for random assignment blocks.

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

SOURCES: Spring 2013 Teacher Survey.

Exhibit J.3. Background characteristics of teachers in the Year 2 principal leadership impact sample, overall and within CLASS and FFT districts, by treatment status

Characteristic	Treatment group	Control group	Estimated difference	p value
All districts				
Year of teaching experience				
Years of experience in district	10.6	10.5	0.2	0.833
Mean number of years	13.2	13.1	0.1	0.930
Three years or fewer (percentage)	12.8	18.0	-5.2*	0.043
Four to 10 years (percentage)	36.1	29.8	6.3*	0.044
Eleven to 20 years (percentage)	29.3	28.9	0.4	0.912
More than 20 years (percentage)	21.9	23.4	-1.5	0.628
Master's degree or higher (percentage)	45.7	44.6	1.1	0.746
CLASS districts				
Year of teaching experience				
Years of experience in district	11.3	10.4	0.9	0.392
Mean number of years	13.6	12.7	0.9	0.393
Three years or fewer (percentage)	11.9	18.7	-6.8*	0.049
Four to 10 years (percentage)	35.4	29.5	5.9	0.122
Eleven to 20 years (percentage)	28.9	31.8	-2.9	0.459
More than 20 years (percentage)	23.7	20.3	3.4	0.396
Master's degree or higher (percentage)	34.5	35.3	-0.8	0.804
FFT districts				
Year of teaching experience				
Years of experience in district	10.0	10.6	-0.6	0.549
Mean number of years	12.7	13.6	-0.9	0.364
Three years or fewer (percentage)	13.6	17.1	-3.5	0.346
Four to 10 years (percentage)	36.7	30.0	6.6	0.180
Eleven to 20 years (percentage)	29.6	26.2	3.4	0.509
More than 20 years (percentage)	20.2	26.4	-6.3	0.191
Master's degree or higher (percentage)	56.5	53.0	3.5	0.568

NOTES: Sample size for all districts = 63 schools and 499 teachers for the treatment group; 63 schools and 524 teachers for the control group. Sample size for CLASS districts = 31 schools and 301 teachers for the treatment group; 32 schools and 313 teachers for the control group. Sample size for FFT districts = 32 schools and 198 teachers for the treatment group; 31 schools and 211 teachers for the control group.

The analyses were based on a two-level regression (teachers within schools) controlling for random assignment blocks.

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

SOURCES: Spring 2014 Teacher Survey.

Exhibit J.4. Background characteristics of principals in the Year 1 principal leadership impact sample, overall and within CLASS and FFT districts, by treatment status

Characteristic	Treatment group	Control group	Estimated difference	p value
All districts				
Years of experience as a principal				
Mean number of years	8.0	9.8	-1.8	0.080
Three years or fewer (percentage)	24.4	16.1	8.3	0.239
Four to 10 years (percentage)	47.5	47.9	-0.3	0.971
Eleven to 20 years (percentage)	22.8	28.6	-5.8	0.450
More than 20 years (percentage)	5.3	7.4	-2.2	0.638
Mean number of years teaching	12.1	13.0	-0.9	0.361
CLASS districts				
Years of experience as a principal				
Mean number of years	6.6	10.7	-4.0*	0.007
Three years or fewer (percentage)	35.5	18.4	17.1	0.117
Four to 10 years (percentage)	38.7	42.3	-3.6	0.795
Eleven to 20 years (percentage)	25.8	26.3	-0.5	0.964
More than 20 years (percentage)	0.0	13.0	-13.0	0.058
Mean number of years teaching	10.5	13.4	-2.9*	0.035
FFT districts				
Years of experience as a principal				
Mean number of years	9.4	9.0	0.4	0.801
Three years or fewer (percentage)	13.7	13.9	-0.3	0.977
Four to 10 years (percentage)	56.0	53.2	2.8	0.832
Eleven to 20 years (percentage)	19.9	30.8	-10.9	0.340
More than 20 years (percentage)	†	†	8.3	0.195
Mean number of years teaching	9.4	9.0	0.4	0.801

NOTES: Sample size for all districts = 61 treatment principals and 62 control principals. Sample size for CLASS = 31 treatment principals and 31 control principals. Sample size for FFT districts = 30 treatment principals and 31 control principals.

The analyses were based on a principal-level regression controlling for random assignment blocks.

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

† Reporting standards not met; in one or more cells, there are too few cases to report

SOURCE: Spring 2013 Principal Survey.

Exhibit J.5. Background characteristics of principals in the Year 2 principal leadership impact sample, overall and within CLASS and FFT districts, by treatment status

Characteristic	Treatment group	Control group	Estimated difference	p value
All districts				
Years of experience as a principal				
Mean number of years	8.0	9.6	-1.5	0.118
Three years or fewer (percentage)	28.8	15.0	13.8	0.072
Four to 10 years (percentage)	38.1	47.5	-9.5	0.318
Eleven to 20 years (percentage)	29.9	30.7	-0.7	0.932
More than 20 years (percentage)	3.2	6.8	-3.6	0.353
Mean number of years teaching	12.5	11.9	0.6	0.562
CLASS districts				
Years of experience as a principal				
Mean number of years	6.9	9.6	-2.8*	0.040
Three years or fewer (percentage)	32.7	24.6	8.1	0.549
Four to 10 years (percentage)	35.5	35.3	0.2	0.990
Eleven to 20 years (percentage)	31.8	33.8	-2.0	0.872
More than 20 years (percentage)	0.0	6.3	-6.3	0.138
Mean number of years teaching	9.5	11.6	-2.1	0.117
FFT districts				
Years of experience as a principal				
Mean number of years	9.2	9.6	-0.4	0.799
Three years or fewer (percentage)	25.0	5.6	19.4*	0.018
Four to 10 years (percentage)	40.6	59.4	-18.8	0.149
Eleven to 20 years (percentage)	28.1	27.7	0.4	0.972
More than 20 years (percentage)	6.3	7.3	-1.1	0.868
Mean number of years teaching	15.3	12.2	3.1*	0.034

NOTES: Sample size for all districts = 61 treatment principals and 58 or 59 control principals. Sample size for CLASS = 29 treatment principals and 29 or 30 control principals. Sample size for FFT districts = 32 treatment principals and 29 control principals. The analyses were based on a principal-level regression controlling for random assignment blocks.

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

SOURCE: Spring 2014 Principal Survey.

Exhibit J.6. Background characteristics of students in Year 1 reading/ELA achievement impact sample, by treatment status

Characteristic	Treatment group	Control group	Estimated difference	p value
Students eligible for free or reduced-price lunch (percentage)	60.3	62.0	-1.7	0.306
Female (percentage)	49.8	49.0	0.9	0.205
Non-White (percentage)	56.3	57.2	-0.9	0.465
English language learners (percentage)	12.5	14.0	-1.6	0.249
Students with disabilities (percentage)	8.8	8.9	-0.1	0.857
2011–12 Student achievement on state assessment in reading (standardized)	0.019	0.029	-0.010	0.735
Grade level (percentage)				
4th grade	23.0	23.5	-0.6	0.856
5th grade	22.5	21.6	0.9	0.770
6th grade	17.8	19.6	-1.8	0.448
7th grade	19.2	18.3	1.0	0.565
8th grade	17.3	16.9	0.5	0.772

NOTES: Sample size = 63 schools, 384 teachers, and 13,134 students for the treatment group; 64 schools, 421 teachers, and 15,358 students for the control group.

The analyses were based on a three-level regression (students within teachers within schools) controlling for random assignment blocks.

None of the differences between the treatment and the control groups are statistically significant at the .05 level (two-tailed).

SOURCES: District Administrative Records.

Exhibit J.7. Background characteristics of students in Year 1 reading/ELA achievement impact sample in CLASS and FFT districts, by treatment status

Characteristic	Treatment group	Control group	Estimated difference	p value
CLASS Districts				
Students eligible for free or reduced-price lunch (percentage)	67.1	68.3	-1.1	0.608
Female (percentage)	50.6	48.9	1.7	0.079
Non-White (percentage)	72.9	71.7	1.2	0.279
English language learners (percentage)	22.9	25.3	-2.4	0.333
Students with disabilities (percentage)	6.0	6.1	-0.1	0.925
2011–12 Student achievement on state assessment in mathematics (standardized)	0.033	-0.005	0.038	0.308
Grade level (percentage)	24.9	25.8	-0.9	0.840
4th grade	24.1	23.1	1.0	0.819
5th grade	15.0	17.4	-2.4	0.293
6th grade	18.1	17.8	0.3	0.882
7th grade	17.8	15.7	2.1	0.335
8th grade	67.1	68.3	-1.1	0.608
FFT districts				
Students eligible for free or reduced-price lunch (percentage)	53.7	56.0	-2.3	0.367
Female (percentage)	49.1	49.0	0.1	0.892
Non-White (percentage)	40.3	43.1	-2.8	0.209
English language learners (percentage)	2.4	3.3	-0.9	0.185
Students with disabilities (percentage)	11.4	11.5	-0.1	0.920
2011–12 Student achievement on state assessment in mathematics (standardized)	0.005	0.066	-0.060	0.210
Grade level (percentage)	21.1	21.4	-0.3	0.947
4th grade	21.0	20.1	0.9	0.849
5th grade	20.6	21.7	-1.2	0.794
6th grade	20.3	18.7	1.6	0.493
7th grade	16.8	18.0	-1.1	0.623
8th grade	53.7	56.0	-2.3	0.367

NOTES: Sample size for CLASS districts = 31 schools, 203 teachers, and 7,402 students for the treatment group; 32 schools, 240 teachers, and 8,447 students for the control group. Sample size for FFT districts = 32 schools, 181 teachers, and 5,732 students for the treatment group; 32 schools, 181 teachers, and 6,911 students for the control group. The analyses were based on a three-level regression (students within teachers within schools) controlling for random assignment blocks.

None of the differences between the treatment and the control groups are statistically significant at the .05 level (two-tailed).

SOURCES: District Administrative Records.

Exhibit J.8. Background characteristics of students in Year 1 mathematics achievement impact sample, by treatment status

Characteristic	Treatment group	Control group	Estimated difference	p value
Students eligible for free or reduced-price lunch (percentage)	60.7	62.5	-1.7	0.295
Female (percentage)	49.3	48.6	0.8	0.265
Non-White (percentage)	56.4	57.4	-0.9	0.410
English language learners (percentage)	14.2	15.3	-1.2	0.422
Students with disabilities (percentage)	8.7	9.2	-0.5	0.398
2011–12 Student achievement on state assessment in mathematics (standardized)	0.021	0.003	0.018	0.627
Grade level (percentage)				
4th grade	23.0	22.5	0.5	0.879
5th grade	21.5	21.9	-0.4	0.895
6th grade	19.8	19.7	0.0	0.996
7th grade	18.8	19.0	-0.2	0.915
8th grade	16.9	16.8	0.1	0.940

NOTES: Sample size = 63 schools, 411 teachers, and 13,967 students for the treatment group; 64 schools, 439 teachers, and 15,907 students for the control group.

The analyses were based on a three-level regression (students within teachers within schools) controlling for random assignment blocks.

None of the differences between the treatment and the control groups are statistically significant at the .05 level (two-tailed).

SOURCE: District Administrative Records.

Exhibit J.9. Background characteristics of students in Year 1 mathematics achievement impact sample in CLASS and FFT districts, by treatment status

Characteristic	Treatment group	Control group	Estimated difference	p value
CLASS districts				
Students eligible for free or reduced-price lunch (percentage)	67.9	68.7	-0.8	0.705
Female (percentage)	49.9	48.6	1.3	0.125
Non-White (percentage)	73.2	72.0	1.2	0.259
English language learners (percentage)	26.2	28.2	-2.0	0.450
Students with disabilities (percentage)	6.4	6.4	0.0	0.985
2011–12 Student achievement on state assessment in mathematics (standardized)	-0.001	-0.068	0.067	0.221
Grade level (percentage)	24.9	24.1	0.8	0.835
4th grade	22.0	22.9	-0.8	0.835
5th grade	19.0	18.7	0.3	0.901
6th grade	18.3	18.8	-0.4	0.846
7th grade	15.8	15.6	0.2	0.942
8th grade	67.9	68.7	-0.8	0.705
FFT districts				
Students eligible for free or reduced-price lunch (percentage)	53.8	56.4	-2.6	0.306
Female (percentage)	48.8	48.8	0.0	1.000
Non-White (percentage)	40.2	43.2	-3.0	0.157
English language learners (percentage)	2.5	3.1	-0.6	0.269
Students with disabilities (percentage)	10.9	12.0	-1.0	0.385
2011–12 Student achievement on state assessment in mathematics (standardized)	0.038	0.065	-0.027	0.587
Grade level (percentage)	21.1	20.9	0.1	0.978
4th grade	21.0	21.0	0.0	1.000
5th grade	20.5	20.8	-0.3	0.952
6th grade	19.3	19.2	0.1	0.958
7th grade	18.1	18.1	0.0	0.988
8th grade	53.8	56.4	-2.6	0.306

NOTES: Sample size for CLASS districts = 31 schools, 232 teachers, and 8,269 students for the treatment group; 32 schools, 257 teachers, and 9,148 students for the control group. Sample size for FFT districts = 32 schools, 179 teachers, and 5,698 students for the treatment group; 32 schools, 182 teachers, and 6,759 students for the control group.

The analyses were based on a three-level regression (students within teachers within schools) controlling for random assignment blocks.

None of the differences between the treatment and the control groups are statistically significant at the .05 level (two-tailed).

SOURCES: District Administrative Records.

Exhibit J.10. Background characteristics of students in Year 2 reading/ELA achievement impact sample, by treatment status

Characteristic	Treatment group	Control group	Estimated difference	p value
Students eligible for free or reduced-price lunch (percentage)	61.2	61.9	-0.7	0.652
Female (percentage)	48.9	49.0	-0.2	0.800
Non-White (percentage)	57.2	56.9	0.3	0.772
English language learners (percentage)	12.8	14.3	-1.5	0.323
Students with disabilities (percentage)	8.2	8.5	-0.4	0.592
2011–12 Student achievement on state assessment in reading (standardized)	0.048	0.080	-0.032	0.367
Grade level (percentage)				
4th grade	20.3	21.2	-1.0	0.764
5th grade	20.9	20.0	0.9	0.787
6th grade	20.4	20.6	-0.1	0.951
7th grade	20.0	18.8	1.2	0.497
8th grade	18.4	19.3	-1.0	0.583

NOTES: Sample size = 63 schools, 374 teachers, and 13,962 students for the treatment group; 63 schools, 394 teachers, and 15,423 students for the control group.

The analyses were based on a three-level regression (students within teachers within schools) controlling for random assignment blocks.

None of the differences between the treatment and the control groups are statistically significant at the .05 level (two-tailed).

SOURCES: District Administrative Records.

Exhibit J.11. Background characteristics of students in Year 2 reading/ELA achievement impact sample in CLASS and FFT districts, by treatment status

Characteristic	Treatment group	Control group	Estimated difference	p value
CLASS districts				
Students eligible for free or reduced-price lunch (percentage)	66.7	66.6	0.1	0.965
Female (percentage)	49.3	49.6	-0.2	0.805
Non-White (percentage)	72.7	71.7	1.0	0.370
English language learners (percentage)	21.4	24.2	-2.8	0.297
Students with disabilities (percentage)	6.1	5.7	0.4	0.470
2011–12 Student achievement on state assessment in mathematics (standardized)	0.044	0.043	0.001	0.991
Grade level (percentage)	20.2	21.6	-1.4	0.738
4th grade	22.1	20.7	1.4	0.736
5th grade	20.1	20.1	0.0	0.998
6th grade	20.0	20.0	0.0	0.990
7th grade	17.5	17.5	0.0	0.991
8th grade	66.7	66.6	0.1	0.965
FFT districts				
Students eligible for free or reduced-price lunch (percentage)	56.0	57.4	-1.4	0.509
Female (percentage)	48.4	48.6	-0.2	0.880
Non-White (percentage)	42.2	42.5	-0.2	0.913
English language learners (percentage)	4.5	4.8	-0.3	0.826
Students with disabilities (percentage)	10.2	11.3	-1.0	0.433
2011–12 Student achievement on state assessment in mathematics (standardized)	0.053	0.109	-0.056	0.208
Grade level (percentage)	20.3	20.9	-0.6	0.908
4th grade	19.6	19.3	0.4	0.942
5th grade	20.7	21.0	-0.3	0.949
6th grade	20.0	17.6	2.4	0.337
7th grade	19.2	21.1	-1.9	0.470
8th grade	56.0	57.4	-1.4	0.509

NOTES: Sample size for CLASS districts = 31 schools, 208 teachers, and 8,059 students for the treatment group; 32 schools, 231 teachers, and 8,997 students for the control group. Sample size for FFT districts = 32 schools, 166 teachers, and 5,903 students for the treatment group; 31 schools, 163 teachers, and 6,426 students for the control group.

The analyses were based on a three-level regression (students within teachers within schools) controlling for random assignment blocks.

None of the differences between the treatment and the control groups are statistically significant at the .05 level (two-tailed).

SOURCES: District Administrative Records.

Exhibit J.12. Background characteristics of students in Year 2 mathematics achievement impact sample, by treatment status

Characteristic	Treatment group	Control group	Estimated difference	p value
Students eligible for free or reduced-price lunch (percentage)	62.0	62.7	-0.6	0.673
Female (percentage)	48.6	49.1	-0.4	0.519
Non-White (percentage)	57.5	57.1	0.4	0.698
English language learners (percentage)	15.1	16.2	-1.0	0.550
Students with disabilities (percentage)	8.5	8.8	-0.3	0.707
2011–12 Student achievement on state assessment in mathematics (standardized)	0.009	0.030	-0.021	0.546
Grade level (percentage)				
4th grade	21.8	22.2	-0.3	0.919
5th grade	22.0	21.8	0.3	0.936
6th grade	19.6	20.0	-0.5	0.847
7th grade	19.4	18.9	0.5	0.780
8th grade	17.2	17.1	0.0	0.982

NOTES: Sample size = 63 schools, 389 teachers, and 14,186 students for the treatment group; 63 schools, 396 teachers, and 15,809 students for the control group.

The analyses were based on a three-level regression (students within teachers within schools) controlling for random assignment blocks.

None of the differences between the treatment and the control groups are statistically significant at the .05 level (two-tailed).

SOURCES: District Administrative Records.

Exhibit J.13. Background characteristics of students in Year 2 mathematics achievement impact sample in CLASS and FFT districts, by treatment status

Characteristic	Treatment group	Control group	Estimated difference	p value
CLASS districts				
Students eligible for free or reduced-price lunch (percentage)	68.0	67.8	0.3	0.892
Female (percentage)	48.8	49.6	-0.8	0.399
Non-White (percentage)	73.3	72.0	1.2	0.240
English language learners (percentage)	25.9	27.6	-1.7	0.546
Students with disabilities (percentage)	6.4	5.9	0.4	0.495
2011–12 Student achievement on state assessment in mathematics (standardized)	-0.009	-0.017	0.008	0.763
Grade level (percentage)	23.1	23.2	0.0	0.995
4th grade	23.6	23.6	0.0	0.995
5th grade	17.9	18.9	-1.0	0.673
6th grade	18.7	18.0	0.7	0.750
7th grade	16.7	16.4	0.3	0.906
8th grade	68.0	67.8	0.3	0.892
FFT districts				
Students eligible for free or reduced-price lunch (percentage)	56.2	57.7	-1.5	0.495
Female (percentage)	48.5	48.5	0.0	1.000
Non-White (percentage)	42.3	42.6	-0.3	0.899
English language learners (percentage)	4.7	5.2	-0.5	0.730
Students with disabilities (percentage)	10.6	11.5	-0.9	0.492
2011–12 Student achievement on state assessment in mathematics (standardized)	0.027	0.084	-0.057	0.222
Grade level (percentage)	20.6	21.2	-0.6	0.903
4th grade	20.5	20.0	0.5	0.924
5th grade	21.2	21.2	0.1	0.991
6th grade	20.1	19.8	0.3	0.918
7th grade	17.6	17.8	-0.2	0.943
8th grade	56.2	57.7	-1.5	0.495

NOTES: Sample size for CLASS districts = 31 schools, 230 teachers, and 8,315 students for the treatment group; 32 schools, 235 teachers, and 8,823 students for the control group. Sample size for FFT districts = 32 schools, 159 teachers, and 5,871 students for the treatment group; 31 schools, 161 teachers, and 6,986 students for the control group.

The analyses were based on a three-level regression (students within teachers within schools) controlling for random assignment blocks.

None of the differences between the treatment and the control groups are statistically significant at the .05 level (two-tailed).

SOURCES: District Administrative Records.

Supporting Exhibits for Classroom Practice Impact Analyses

Exhibit J.14. Average CLASS and FFT overall scores, based on coding of video-recorded lessons by study team, overall and within CLASS and FFT districts, by treatment status, Year 2

Classroom practice measure	Treatment group mean	Control group mean	Estimated difference	Standard error	Effect size	p value
All districts						
Mean CLASS overall score	4.50	4.39	0.11*	0.04	0.17	0.006
Mean FFT overall score	2.65	2.63	0.02	0.02	0.04	0.418
CLASS districts						
Mean CLASS overall score	4.64	4.32	0.31†	0.06	0.46	0.000
Mean FFT overall score	2.67	2.61	0.06	0.03	0.14	0.069
FFT districts						
Mean CLASS overall score	4.37	4.44	-0.07†	0.06	-0.09	0.287
Mean FFT overall score	2.63	2.64	0.00	0.04	-0.01	0.916

NOTES: Sample size for all districts = 63 schools, 434 teachers, and 668 lessons for the treatment group; 63 schools, 517 teachers, and 793 lessons for the control group. Sample size for CLASS districts = 63 schools, 238 teachers, and 360 lessons for the treatment group; 63 schools, 306 teachers, and 462 lessons for the control group. Sample size for FFT districts = 63 schools, 211 teachers, and 308 lessons for the treatment group; 63 schools, 232 teachers, and 331 lessons for the control group.

The analyses were based on a three-level regression (lessons within teachers within schools) controlling for random assignment blocks and teacher background characteristics.

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

† The difference between CLASS districts and FFT districts in the estimated difference is statistically significant at the .05 level (two-tailed).

SOURCE: Spring 2014 Classroom Videos.

Exhibit J.15. Average CLASS and FFT domain scores, based on coding of video-recorded lessons by study team, by treatment status, Year 2

Classroom practice measure	Treatment group mean	Control group mean	Estimated difference	Standard error	Effect size	p value
CLASS domains						
Emotional support	4.32	4.11	0.20*	0.05	0.21	0.000
Classroom organization	6.11	6.08	0.03	0.05	0.04	0.559
Instructional support	3.51	3.41	0.11*	0.05	0.12	0.034
Student engagement	5.15	5.02	0.13*	0.05	0.15	0.007
FFT domains						
Classroom environment	2.84	2.84	0.00	0.03	-0.01	0.865
Instruction	2.46	2.42	0.04	0.03	0.08	0.122

NOTES: Sample size = 63 schools, 434 teachers, and 668 lessons for the treatment group; 63 schools, 517 teachers, and 793 lessons for the control group.

The analyses were based on a three-level regression (lessons within teachers within schools) controlling for random assignment blocks and teacher background characteristics.

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

SOURCE: Spring 2014 Classroom Videos.

Exhibit J.16. Average CLASS and FFT domain scores in CLASS and FFT districts, based on coding of video-recorded lessons by study team, by treatment status, Year 2

Classroom Practice Measure	Treatment group mean	Control group mean	Estimated difference	Standard error	Effect size	p value
CLASS districts						
<u>CLASS domains</u>						
Emotional support	4.44	3.91	0.52*	0.07	0.56	0.000
Classroom organization	6.12	5.98	0.14*	0.07	0.19	0.039
Instructional support	3.72	3.41	0.31*	0.07	0.37	0.000
Student engagement	5.34	5.10	0.24*	0.06	0.28	0.000
<u>FFT domains</u>						
Classroom environment	2.85	2.85	0.00	0.03	0.00	0.959
Instruction	2.49	2.38	0.11*	0.04	0.22	0.003
FFT districts						
<u>CLASS domains</u>						
Emotional support	4.20	4.29	-0.09	0.08	-0.10	0.275
Classroom organization	6.10	6.17	-0.07	0.07	-0.10	0.311
Instructional support	3.31	3.38	-0.07	0.07	-0.08	0.340
Student engagement	4.97	4.94	0.03	0.07	0.03	0.678
<u>FFT domains</u>						
Classroom environment	2.82	2.82	0.00	0.04	0.00	0.985
Instruction	2.44	2.45	-0.01	0.05	-0.01	0.875

NOTES: Sample size for CLASS districts =30 schools, 238 teachers, and 360 lessons for the treatment group; 32 schools, 306 teachers, and 462 lessons for the control group. Sample size for FFT districts = 31 schools, 211 teachers, and 308 lessons for the treatment group; 31 schools, 232 teachers, and 331 lessons for the control group.

The analyses were based on a three-level regression (lessons within teachers within schools) controlling for random assignment blocks and teacher background characteristics.

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

SOURCE: Spring 2014 Classroom Videos.

Exhibit J.17. Average CLASS and FFT overall scores, based on coding of video-recorded lessons by study team, by treatment status and district, Year 2

District ID and assigned classroom observation system for intervention	Treatment group mean	Control group mean	Estimated difference	Standard error	p value
Mean CLASS overall score					
1	CLASS	4.53	4.39	0.13	0.14
2	CLASS	4.49	4.27	0.22*	0.09
3	CLASS	4.76	4.20	0.56*	0.11
4	CLASS	4.81	4.55	0.26*	0.12
5	FFT	4.38	4.44	-0.07	0.14
6	FFT	4.31	4.44	-0.13	0.09
7	FFT	4.28	4.25	0.02	0.15
8	FFT	4.57	4.69	-0.12	0.12
Chi square			30.88*		0.000
Mean FFT overall score					
1	CLASS	2.66	2.66	0.00	0.08
2	CLASS	2.74	2.67	0.07	0.05
3	CLASS	2.57	2.51	0.06	0.07
4	CLASS	2.68	2.64	0.04	0.07
5	FFT	2.81	2.75	0.06	0.08
6	FFT	2.67	2.71	-0.04	0.06
7	FFT	2.68	2.64	0.04	0.09
8	FFT	2.37	2.44	-0.07	0.08
Chi square			4.29		0.746

NOTES Sample size = 61 schools, 434 teachers, and 668 lessons for the treatment group; 63 schools, 517 teachers, and 793 lessons for the control group across the eight districts.

The analyses were based on a three-level regression (lessons within teachers within schools) controlling for random assignment blocks and teacher background characteristics.

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

SOURCE: Spring 2014 Classroom Videos.

Exhibit J.18. Average CLASS and FFT overall scores without covariate adjustment, based on coding of video-recorded lessons by study team, overall and within CLASS and FFT districts, by treatment status, Year 2

Classroom practice measure	Treatment group mean	Control group mean	Estimated difference	Standard error	Effect size	p value
All districts						
Mean CLASS overall score	4.50	4.38	0.12*	0.04	0.177	0.003
Mean FFT overall score	2.65	2.62	0.03	0.02	0.062	0.263
CLASS districts						
Mean CLASS overall score	4.64	4.31	0.32*	0.06	0.467	0.000
Mean FFT overall score	2.67	2.60	0.06*	0.03	0.153	0.040
FFT districts						
Mean CLASS overall score	4.37	4.44	-0.07	0.06	-0.101	0.253
Mean FFT overall score	2.63	2.64	0.00	0.04	-0.010	0.901

NOTES: Sample size for all districts = 61 schools, 434 teachers, and 668 lessons for the treatment group; 63 schools, 517 teachers, and 793 lessons for the control group across the eight districts. Sample size for CLASS districts =30 schools, 238 teachers, and 360 lessons for the treatment group; 32 schools, 306 teachers, and 462 lessons for the control group. Sample size for FFT districts = 31 schools, 211 teachers, and 308 lessons for the treatment group; 31 schools, 232 teachers, and 331 lessons for the control group.

The analyses were based on a three-level regression (lessons within teachers within schools) controlling for random assignment blocks.

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

SOURCE: Spring 2014 Classroom Videos.

Exhibit J.19. Average CLASS and FFT overall scores, based on coding of video-recorded lessons by study team, overall and within CLASS and FFT districts (excluding District 3), by treatment status, Year 2

Classroom practice measure	Treatment group mean	Control group mean	Estimated difference	Standard error	Effect size	p value
All districts (excluding District 3)						
Mean CLASS overall score	4.46	4.42	0.04	0.05	0.06	0.358
Mean FFT overall score	2.66	2.65	0.01	0.03	0.03	0.619
CLASS districts						
Mean CLASS overall score	4.58	4.38	0.21*	0.07	0.29	0.002
Mean FFT overall score	2.71	2.66	0.05	0.04	0.12	0.160
FFT districts						
Mean CLASS overall score	4.37	4.44	-0.07	0.06	-0.09	0.287
Mean FFT overall score	2.63	2.64	0.00	0.04	-0.01	0.916

NOTES: Sample size for all districts = 54 schools, 383 teachers, and 596 lessons for the treatment group; 54 schools, 419 teachers, and 654 lessons for the control group across the eight districts. Sample size for CLASS districts =22 schools, 187 teachers, and 288 lessons for the treatment group; 23 schools, 208 teachers, and 323 lessons for the control group. Sample size for FFT districts = 31 schools, 211 teachers, and 308 lessons for the treatment group; 31 schools, 232 teachers, and 331 lessons for the control group. The analyses were based on a three-level regression (lessons within teachers within schools) controlling for random assignment blocks and teacher background characteristics.

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

SOURCE: Spring 2014 Classroom Videos.

Supporting Exhibits for Principal Leadership Impact Analyses

Exhibit J.20. Average ratings of principal instructional leadership and teacher-principal trust, by treatment status, and year

Principal leadership measure	Treatment group mean	Control group mean	Estimated difference	Standard error	Effect size	p value
Year 1						
Instructional leadership	3.27	3.19	0.08	0.08	0.10	0.332
Teacher-principal trust	3.18	2.96	0.22*	0.08	0.25	0.006
Year 2						
Instructional leadership	3.35	3.21	0.14*	0.07	0.19	0.045
Teacher-principal trust	3.19	3.03	0.15*	0.08	0.19	0.048

NOTES: Sample size for Year 1 = 63 schools and 524 or 525 teachers for the treatment group; 64 schools and 557 teachers for the control group. Sample size for Year 2 = 63 schools and 499 teachers for the treatment group; 63 schools and 522 or 523 teachers for the control group.

The analyses were based on a two-level regression (teachers within schools) controlling for random assignment blocks and teacher background characteristics.

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

SOURCES: Spring 2013 and Spring 2014 Teacher Surveys.

Exhibit J.21. Average ratings of principal instructional leadership and teacher-principal trust in CLASS and FFT districts, by treatment status and year

Principal leadership measure	Treatment group mean	Control group mean	Estimated difference	Standard error	Effect size	p value
Year 1						
CLASS districts						
Instructional leadership	3.41	3.30	0.11	0.10	0.14	0.266
Teacher-principal trust	3.23	3.05	0.17	0.11	0.20	0.109
FFT districts						
Instructional leadership	3.13	3.09	0.04	0.14	0.06	0.744
Teacher-principal trust	3.14	2.90	0.25*	0.12	0.29	0.041
Year 2						
CLASS districts						
Instructional leadership	3.36	3.33	0.03	0.09	0.04	0.766
Teacher-principal trust	3.11	3.14	-0.03†	0.11	-0.03	0.801
FFT districts						
Instructional leadership	3.33	3.09	0.24*	0.10	0.34	0.014
Teacher-principal trust	3.26	2.93	0.33*†	0.10	0.43	0.001

Notes: Year 1 sample size for CLASS districts = 31 schools and 307 teachers for the treatment group; 32 schools and 328 teachers for the control group. Year 1 sample size for FFT districts = 32 schools and 217 or 218 teachers for the treatment group; 32 schools and 229 teachers for the control group. Year 2 sample size for CLASS districts = 31 schools and 301 teachers for the treatment group; 32 schools and 312 or 313 teachers for the control group. Year 2 sample size for FFT districts = 32 schools and 198 teachers for the treatment group; 31 schools and 210 teachers for the control group.

The analyses were based on a two-level regression (teachers within schools) controlling for random assignment blocks and teacher background characteristics.

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

† The difference between CLASS districts and FFT districts in the estimated difference is statistically significant at the .05 level (two-tailed).

SOURCES: Spring 2013 and Spring 2014 Teacher Surveys

Exhibit J.22. Average rating of principal instructional leadership, by treatment status, district, and year

District number and assigned classroom observation system for intervention	Treatment group mean	Control group mean	Estimated difference	Standard error	p value
Year 1					
1	CLASS	2.92	3.14	-0.22	0.29
2	CLASS	3.73	3.40	0.34	0.19
3	CLASS	3.38	3.48	-0.10	0.21
4	CLASS	3.26	2.97	0.29	0.25
5	FFT	3.08	3.15	-0.06	0.28
6	FFT	2.82	3.06	-0.24	0.19
7	FFT	3.38	3.10	0.28	0.26
8	FFT	3.40	3.04	0.36	0.25
Chi square			9.26		0.235
Year 2					
1	CLASS	2.94	3.30	-0.36	0.24
2	CLASS	3.65	3.31	0.35*	0.15
3	CLASS	3.32	3.50	-0.18	0.17
4	CLASS	3.24	3.17	0.07	0.20
5	FFT	3.39	2.96	0.43	0.25
6	FFT	3.14	2.98	0.16	0.16
7	FFT	3.49	3.38	0.11	0.22
8	FFT	3.44	3.03	0.41*	0.20
Chi square			16.85*		0.018

NOTES: Sample size for Year 1 = 63 schools and 524 or 525 teachers for the treatment group; 64 schools and 557 teachers for the control group. Sample size for Year 2 = 63 schools and 499 teachers for the treatment group; 63 schools and 522 or 523 teachers for the control group.

The analyses were based on a two-level regression (teachers within schools) controlling for random assignment blocks and teacher background characteristics.

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

SOURCES: Spring 2013 and Spring 2014 Teacher Surveys

Exhibit J.23. Average rating of teacher-principal trust, by treatment status, district, and year

District number and assigned classroom observation system for intervention	Treatment group mean	Control group mean	Estimated difference	Standard error	p value
Year 1					
1 CLASS	2.79	3.16	-0.37	0.28	0.197
2 CLASS	3.49	3.12	0.37*	0.18	0.041
3 CLASS	3.17	3.14	0.03	0.20	0.871
4 CLASS	3.18	2.65	0.53*	0.25	0.031
5 FFT	3.14	3.03	0.11	0.27	0.681
6 FFT	2.82	2.97	-0.15	0.19	0.439
7 FFT	3.42	2.83	0.59*	0.26	0.023
8 FFT	3.39	2.72	0.67*	0.24	0.005
Chi square			13.41		0.063
Year 2					
1 CLASS	2.69	3.34	-0.65*	0.27	0.016
2 CLASS	3.41	2.98	0.43*	0.17	0.013
3 CLASS	2.97	3.34	-0.37	0.19	0.051
4 CLASS	3.14	2.96	0.18	0.23	0.441
5 FFT	3.43	2.90	0.53	0.29	0.065
6 FFT	3.05	2.86	0.19	0.18	0.290
7 FFT	3.36	3.10	0.27	0.25	0.288
8 FFT	3.38	2.90	0.48*	0.23	0.039
Chi square			22.98*		0.002

NOTES: Sample size for Year 1 = 63 schools and 524 or 525 teachers for the treatment group; 64 schools and 557 teachers for the control group. Sample size for Year 2 = 63 schools and 499 teachers for the treatment group; 63 schools and 522 or 523 teachers for the control group.

The analyses were based on a two-level regression (teachers within schools) controlling for random assignment blocks and teacher background characteristics.

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

SOURCES: Spring 2013 and Spring 2014 Teacher Surveys.

Exhibit J.24. Average ratings of principal instructional leadership and teacher-principal trust without covariate adjustment, by treatment status and year

Principal leadership measure	Treatment group mean	Control group mean	Estimated difference	Standard error	Effect size	p value
Year 1						
Instructional leadership	3.27	3.19	0.07	0.08	0.09	0.379
Teacher-principal trust	3.18	2.97	0.22*	0.08	0.25	0.007
Year 2						
Instructional leadership	3.35	3.22	0.13	0.07	0.18	0.063
Teacher-principal trust	3.19	3.04	0.15	0.08	0.18	0.060

NOTES: Sample size for Year 1 = 63 schools and 524 or 525 teachers for the treatment group; 64 schools and 557 teachers for the control group. Sample size for Year 2 = 63 schools and 499 teachers for the treatment group; 63 schools and 522 or 523 teachers for the control group.

The analyses were based on a two-level regression (teachers within schools) controlling for random assignment blocks.

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

SOURCES: Spring 2013 and Spring 2014 Teacher Surveys.

Exhibit J.25. Average ratings of principal instructional leadership and teacher-principal trust without covariate adjustment in CLASS and FFT districts, by treatment status and year

Principal leadership measure	Treatment group mean	Control group mean	Estimated difference	Standard error	Effect size	p value
Year 1						
CLASS districts						
Instructional leadership	3.41	3.29	0.11	0.10	0.142	0.247
Teacher-principal trust	3.23	3.05	0.18	0.11	0.202	0.099
FFT districts						
Instructional leadership	3.13	3.09	0.04	0.13	0.053	0.754
Teacher-principal trust	3.14	2.89	0.25*	0.12	0.297	0.033
Year 2						
CLASS districts						
Instructional leadership	3.36	3.34	0.02	0.09	0.03	0.828
Teacher-principal trust	3.11	3.15	-0.03	0.11	-0.04	0.759
FFT districts						
Instructional leadership	3.33	3.10	0.23*	0.10	0.33	0.023
Teacher-principal trust	3.26	2.94	0.32*	0.10	0.43	0.002

NOTES: Year 1 sample size for CLASS districts = 31 schools and 307 teachers for the treatment group; 32 schools and 328 teachers for the control group. Year 1 sample size for FFT districts = 32 schools and 217 or 218 teachers for the treatment group; 32 schools and 229 teachers for the control group. Year 2 sample size for CLASS districts = 31 schools and 301 teachers for the treatment group; 32 schools and 312 or 313 teachers for the control group. Year 2 sample size for FFT districts = 32 schools and 198 teachers for the treatment group; 31 schools and 210 teachers for the control group.

The analyses were based on a two-level regression (teachers within schools) controlling for random assignment blocks.

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

SOURCES: Spring 2013 and Spring 2014 Teacher Surveys.

Supporting Exhibits for Student Achievement Impact Analyses

**Exhibit J.26. Average reading/ELA and mathematics achievement,
by treatment status and year**

Student achievement measure	Treatment group mean	Control group mean	Estimated difference	Standard error	p value
Year 1					
Reading	-0.003	-0.013	0.010	0.018	0.583
Mathematics	0.046	-0.007	0.053*	0.019	0.005
Year 2					
Reading	0.012	-0.012	0.024	0.026	0.345
Mathematics	0.029	-0.029	0.058	0.030	0.055

NOTES: Year 1 sample size for reading = 63 schools, 384 teachers, and 13,134 students for the treatment group; 64 schools, 421 teachers, and 15,358 students for the control group. Year 1 sample size for mathematics = 63 schools, 411 teachers, and 13,967 students for the treatment group; 64 schools, 439 teachers, and 15,907 students for the control group. Year 2 sample size for reading = 63 schools, 374 teachers, and 13,962 students for the treatment group; 63 schools, 394 teachers, and 15,423 students for the control group. Year 2 sample size for mathematics = 63 schools, 389 teachers, and 14,186 students for the treatment group; 63 schools, 396 teachers, and 15,809 students for the control group.

The analyses were based on a three-level regression (students within teachers within schools) controlling for random assignment blocks and student background characteristics.

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

SOURCE: District Administrative Records.

Exhibit J.27. Average reading/ELA and mathematics achievement in CLASS and FFT districts, by treatment status and year

Student achievement measure	Treatment group mean	Control group mean	Estimated difference	Standard error	p value
Year 1					
CLASS districts					
Reading	0.004	0.011	-0.008	0.023	0.730
Mathematics	0.048	-0.001	0.049	0.025	0.0502
FFT districts					
Reading	-0.010	-0.035	0.025	0.028	0.369
Mathematics	0.044	-0.015	0.058*	0.027	0.029
Year 2					
CLASS districts					
Reading	-0.014	0.013	-0.027†	0.032	0.406
Mathematics	0.021	-0.028	0.050	0.041	0.227
FFT districts					
Reading	0.037	-0.039	0.076†	0.041	0.061
Mathematics	0.037	-0.028	0.064	0.043	0.137

NOTES: Year 1 sample size for reading in CLASS districts = 31 schools, 203 teachers, and 7,402 students for the treatment group; 32 schools, 240 teachers, and 8,447 students for the control group. Year 1 sample size for mathematics in CLASS districts = 31 schools, 232 teachers, and 8,269 students for the treatment group; 32 schools, 257 teachers, and 9,148 students for the control group. Year 1 sample size for reading in FFT districts = 32 schools, 181 teachers, and 5,732 students for the treatment group; 32 schools, 181 teachers, and 6,911 students for the control group. Year 1 sample size for mathematics in FFT districts = 32 schools, 179 teachers, and 5,698 students for the treatment group; 32 schools, 182 teachers, and 6,759 students for the control group. Year 2 sample size for reading in CLASS districts = 31 schools, 208 teachers, and 8,059 students for the treatment group; 32 schools, 231 teachers, and 8,997 students for the control group. Year 2 sample size for mathematics in CLASS districts = 31 schools, 230 teachers, and 8,315 students for the treatment group; 32 schools, 235 teachers, and 8,823 students for the control group. Year 2 sample size for reading in FFT districts = 32 schools, 166 teachers, and 5,903 students for the treatment group; 31 schools, 163 teachers, and 6,426 students for the control group. Year 2 sample size for mathematics in FFT districts = 32 schools, 159 teachers, and 5,871 students for the treatment group; 31 schools, 161 teachers, and 6,986 students for the control group.

The analyses were based on a three-level regression (students within teachers within schools) controlling for random assignment blocks and student background characteristics.

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

† The difference between CLASS districts and FFT districts in the estimated difference is statistically significant at the .05 level (two-tailed).

SOURCE: District Administrative Records.

Exhibit J.28. Average reading/ELA achievement, by treatment status, district, and year

District number and assigned classroom observation system for intervention	Treatment group mean	Control group mean	Estimated difference	Standard error	p value	
Year 1						
1	CLASS	-0.066	-0.070	0.005	0.064	0.941
2	CLASS	0.037	0.108	-0.071	0.040	0.080
3	CLASS	0.061	0.036	0.026	0.046	0.573
4	CLASS	-0.086	-0.109	0.023	0.053	0.660
5	FFT	0.057	0.013	0.044	0.059	0.455
6	FFT	0.015	-0.040	0.056	0.042	0.186
7	FFT	-0.063	-0.068	0.005	0.058	0.936
8	FFT	-0.040	-0.036	-0.003	0.052	0.946
Chi square			6.02		0.537	
Year 2						
1	CLASS	-0.076	0.043	-0.119	0.092	0.193
2	CLASS	-0.003	0.025	-0.028	0.059	0.637
3	CLASS	0.031	0.041	-0.010	0.067	0.886
4	CLASS	-0.049	-0.074	0.025	0.074	0.735
5	FFT	0.072	-0.046	0.118	0.094	0.211
6	FFT	0.033	-0.020	0.053	0.058	0.360
7	FFT	-0.005	-0.056	0.052	0.084	0.538
8	FFT	0.066	-0.039	0.105	0.077	0.170
Chi square			5.93		0.548	

NOTES: Year 1 sample = 63 schools, 384 teachers, and 13,134 students for the treatment group; 64 schools, 421 teachers, and 15,358 students for the control group. Year 2 sample size = 63 schools, 374 teachers, and 13,962 students for the treatment group; 63 schools, 394 teachers, and 15,423 students for the control group.

The analyses were based on a three-level regression (students within teachers within schools) controlling for random assignment blocks and student background characteristics.

None of the differences between the treatment and the control groups are statistically significant at the .05 level (two-tailed).

SOURCE: District Administrative Records.

Exhibit J.29. Average mathematics achievement, by treatment status, district, and year

District number and assigned classroom observation system for intervention	Treatment group mean	Control group mean	Estimated difference	Standard error	p value
Year 1					
1 CLASS	-0.046	-0.061	0.015	0.064	0.815
2 CLASS	0.062	-0.006	0.068	0.040	0.085
3 CLASS	0.093	0.002	0.091	0.046	0.051
4 CLASS	0.033	0.057	-0.024	0.051	0.631
5 FFT	0.099	0.060	0.039	0.062	0.531
6 FFT	0.033	-0.038	0.070	0.043	0.106
7 FFT	0.010	-0.064	0.073	0.066	0.267
8 FFT	0.061	0.033	0.028	0.056	0.619
Chi square			3.96		0.785
Year 2					
1 CLASS	-0.052	-0.032	-0.020	0.105	0.852
2 CLASS	0.035	-0.088	0.123	0.065	0.057
3 CLASS	0.096	0.059	0.036	0.076	0.634
4 CLASS	-0.053	-0.039	-0.015	0.088	0.865
5 FFT	0.044	0.015	0.030	0.111	0.790
6 FFT	0.025	-0.057	0.082	0.068	0.227
7 FFT	0.014	0.007	0.006	0.103	0.951
8 FFT	0.078	-0.061	0.139	0.094	0.138
Chi square			3.51		0.834

NOTES: Year 1 sample size = 63 schools, 411 teachers, and 13,967 students for the treatment group; 64 schools, 439 teachers, and 15,907 students for the control group. Year 2 sample size = 63 schools, 389 teachers, and 14,186 students for the treatment group; 63 schools, 396 teachers, and 15,809 students for the control group.

The analyses were based on a three-level regression (students within teachers within schools) controlling for random assignment blocks and student background characteristics.

None of the differences between the treatment and the control groups are statistically significant at the .05 level (two-tailed).

SOURCE: District Administrative Records.

Exhibit J.30. Average reading/ELA and mathematics achievement without covariate adjustment, by treatment status and year

Student achievement measure	Treatment group mean	Control group mean	Estimated difference	Standard error	p value
Year 1					
Reading	-0.003	-0.007	0.004	0.032	0.893
Mathematics	0.046	-0.023	0.068	0.038	0.073
Year 2					
Reading	0.012	-0.013	0.025	0.034	0.469
Mathematics	0.029	-0.039	0.068	0.043	0.108

NOTES: Year 1 sample size for reading = 63 schools, 384 teachers, and 13,134 students for the treatment group; 64 schools, 421 teachers, and 15,358 students for the control group. Year 1 sample size for mathematics = 63 schools, 411 teachers, and 13,967 students for the treatment group; 64 schools, 439 teachers, and 15,907 students for the control group. Year 2 sample size for reading = 63 schools, 374 teachers, and 13,962 students for the treatment group; 63 schools, 394 teachers, and 15,423 students for the control group. Year 2 sample size for mathematics = 63 schools, 389 teachers, and 14,186 students for the treatment group; 63 schools, 396 teachers, and 15,809 students for the control group.

The analyses were based on a three-level regression (students within teachers within schools) controlling for random assignment blocks.

None of the differences between the treatment and the control groups are statistically significant at the .05 level (two-tailed).

SOURCE: District Administrative Records.

Exhibit J.31. Average reading/ELA and mathematics achievement without covariate adjustment in CLASS and FFT districts, by treatment status and year

Student achievement measure	Treatment group mean	Control group mean	Estimated difference	Standard error	p value
<u>Year 1</u>					
CLASS districts					
Reading	0.004	-0.024	0.027	0.034	0.417
Mathematics	0.048	-0.054	0.102*	0.050	0.040
FFT districts					
Reading	-0.010	0.012	-0.022	0.054	0.684
Mathematics	0.044	0.008	0.035	0.058	0.543
<u>Year 2</u>					
CLASS districts					
Reading	-0.014	0.002	-0.016	0.041	0.699
Mathematics	0.021	-0.068	0.090	0.060	0.134
FFT districts					
Reading	0.037	-0.027	0.064	0.054	0.234
Mathematics	0.037	-0.011	0.048	0.059	0.417

NOTES: Year 1 sample size for reading in CLASS districts = 31 schools, 203 teachers, and 7,402 students for the treatment group; 32 schools, 240 teachers, and 8,447 students for the control group. Year 1 sample size for mathematics in CLASS districts = 31 schools, 232 teachers, and 8,269 students for the treatment group; 32 schools, 257 teachers, and 9,148 students for the control group. Year 1 sample size for reading in FFT districts = 32 schools, 181 teachers, and 5,732 students for the treatment group; 32 schools, 181 teachers, and 6,911 students for the control group. Year 1 sample size for mathematics in FFT districts = 32 schools, 179 teachers, and 5,698 students for the treatment group; 32 schools, 182 teachers, and 6,759 students for the control group. Year 2 sample size for reading in CLASS districts = 31 schools, 208 teachers, and 8,059 students for the treatment group; 32 schools, 231 teachers, and 8,997 students for the control group. Year 2 sample size for mathematics in CLASS districts = 31 schools, 230 teachers, and 8,315 students for the treatment group; 32 schools, 235 teachers, and 8,823 students for the control group. Year 2 sample size for reading in FFT districts = 32 schools, 166 teachers, and 5,903 students for the treatment group; 31 schools, 163 teachers, and 6,426 students for the control group. Year 2 sample size for mathematics in FFT districts = 32 schools, 159 teachers, and 5,871 students for the treatment group; 31 schools, 161 teachers, and 6986 students for the control group.

The analyses were based on a three-level regression (students within teachers within schools) controlling for random assignment blocks.

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

SOURCE: District Administrative Records.

Exhibit J.32. Average reading/ELA and mathematics achievement adjusted for prior achievement in both reading/ELA and mathematics, by treatment status and year

Student achievement measure	Treatment group mean	Control group mean	Estimated difference	Standard error	p value
Year 1					
Reading	-0.003	-0.012	0.009	0.018	0.618
Mathematics	0.046	-0.010	0.056*	0.019	0.003
Year 2					
Reading	0.012	-0.014	0.026	0.025	0.302
Mathematics	0.029	-0.031	0.060*	0.030	0.045

NOTES: Year 1 sample size for reading = 63 schools, 384 teachers, and 13,134 students for the treatment group; 64 schools, 421 teachers, and 15,358 students for the control group. Year 1 sample size for mathematics = 63 schools, 411 teachers, and 13,967 students for the treatment group; 64 schools, 439 teachers, and 15,907 students for the control group. Year 2 sample size for reading = 63 schools, 374 teachers, and 13,962 students for the treatment group; 63 schools, 394 teachers, and 15,423 students for the control group. Year 2 sample size for mathematics = 63 schools, 389 teachers, and 14,186 students for the treatment group; 63 schools, 396 teachers, and 15,809 students for the control group.

The analyses were based on a three-level regression (students within teachers within schools) controlling for random assignment blocks and student background characteristics.

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

SOURCE: District Administrative Records.

Exhibit J.33. Average reading/ELA and mathematics achievement adjusted for prior achievement in both reading/ELA and mathematics in CLASS and FFT districts, by treatment status and year

Student achievement measure	Treatment group mean	Control group mean	Estimated difference	Standard error	p value
<u>Year 1</u>					
CLASS districts					
Reading	0.004	0.014	-0.011	0.023	0.638
Mathematics	0.048	0.001	0.047	0.025	0.059
FFT districts					
Reading	-0.010	-0.036	0.026	0.027	0.330
Mathematics	0.044	-0.022	0.065*	0.027	0.016
<u>Year 2</u>					
CLASS districts					
Reading	-0.014	0.013	-0.027	0.031	0.390
Mathematics	0.021	-0.031	0.053	0.041	0.202
FFT districts					
Reading	0.037	-0.045	0.082*	0.041	0.044
Mathematics	0.037	-0.029	0.066	0.043	0.122

NOTES: Year 1 sample size for reading in CLASS districts = 31 schools, 203 teachers, and 7,402 students for the treatment group; 32 schools, 240 teachers, and 8,447 students for the control group. Year 1 sample size for mathematics in CLASS districts = 31 schools, 232 teachers, and 8,269 students for the treatment group; 32 schools, 257 teachers, and 9,148 students for the control group. Year 1 sample size for reading in FFT districts = 32 schools, 181 teachers, and 5,732 students for the treatment group; 32 schools, 181 teachers, and 6,911 students for the control group. Year 1 sample size for mathematics in FFT districts = 32 schools, 179 teachers, and 5,698 students for the treatment group; 32 schools, 182 teachers, and 6,759 students for the control group. Year 2 sample size for reading in CLASS districts = 31 schools, 208 teachers, and 8,059 students for the treatment group; 32 schools, 231 teachers, and 8,997 students for the control group. Year 2 sample size for mathematics in CLASS districts = 31 schools, 230 teachers, and 8,315 students for the treatment group; 32 schools, 235 teachers, and 8,823 students for the control group. Year 2 sample size for reading in FFT districts = 32 schools, 166 teachers, and 5,903 students for the treatment group; 31 schools, 163 teachers, and 6,426 students for the control group. Year 2 sample size for mathematics in FFT districts = 32 schools, 159 teachers, and 5,871 students for the treatment group; 31 schools, 161 teachers, and 6,986 students for the control group.

The analyses were based on a three-level regression (students within teachers within schools) controlling for random assignment blocks and student background characteristics.

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

SOURCE: District Administrative Records.

Supporting Exhibits for Moderator Analyses

Exhibit J.34. Differential impact of intervention on CLASS and FFT overall scores for teachers with different probationary status, teachers with different prior value-added scores, and teachers at different school levels, Year 2

	CLASS Overall score			FFT Overall score		
	Estimate	Standard error	p value	Estimate	Standard error	p value
Probationary teachers	0.14	0.07	0.054	0.03	0.05	0.558
Nonprobationary teachers	0.11*	0.05	0.033	0.02	0.03	0.536
<i>Difference in effect</i>	0.04	0.09	0.679	0.01	0.06	0.883
Prior value-added score of 0.2	-0.09	0.10	0.345	-0.03	0.06	0.633
Prior value-added score of -0.2	0.33*	0.10	0.001	0.07	0.06	0.266
<i>Difference in effect</i>	-0.42*	0.18	0.021	-0.10	0.11	0.380
Middle schools	0.00	0.09	0.999	0.04	0.04	0.268
Elementary schools	0.17*	0.06	0.003	0.00	0.06	0.958
<i>Difference in effect</i>	-0.17	0.10	0.094	-0.05	0.07	0.515

NOTES: Sample size = 61 schools, 434 teachers, and 668 lessons for the treatment group; 63 schools, 517 teachers, and 793 lessons for the control group.

The analyses were based on a three-level regression (lessons within teachers within schools) controlling for random assignment blocks and teacher background characteristics. The estimates of differential impact represent the differences in impact between probationary and nonprobationary teachers, between teachers whose prior value-added differed by one standard deviation, and between middle school and elementary school teachers, respectively.

* Estimate is statistically significant at the .05 level (two-tailed).

SOURCES: Spring 2014 Classroom Videos; Spring 2014 Teacher Survey; AIR Value-Added System.

Exhibit J.35. Differential impact of intervention on principal instructional leadership and teacher-principal trust for middle school principals and elementary school principals, by year

	Instructional leadership			Teacher-principal trust		
	Estimate	Standard error	p value	Estimate	Standard error	p value
Year 1						
Middle schools	-0.02	0.17	0.902	0.17	0.16	0.283
Elementary schools	0.06	0.09	0.492	0.19*	0.09	0.031
<i>Difference in effect</i>	-0.09	0.20	0.663	-0.02	0.18	0.897
Year 2						
Middle schools	0.09	0.14	0.541	0.20	0.16	0.203
Elementary schools	0.14	0.08	0.094	0.12	0.09	0.186
<i>Difference in effect</i>	-0.05	0.17	0.775	0.08	0.18	0.656

NOTES: Sample size for Year 1 = 63 schools and 524 or 525 teachers for the treatment group; 64 schools and 557 teachers for the control group. Sample size for Year 2 = 63 schools and 499 teachers for the treatment group; 63 schools and 522 or 523 teachers for the control group.

The analyses were based on a two-level regression (teachers within schools) controlling for random assignment blocks and teacher background characteristics. The estimates of differential impact represent the differences in impact between middle school principals and elementary school principals.

None of the estimates are statistically significant at the .05 level (two-tailed).

SOURCES: Spring 2013 and Spring 2014 Teacher Surveys.

Exhibit J.36. Differential impact of intervention on student achievement in reading/ELA and mathematics, for teachers with different probationary status, teachers with different prior value-added, and teachers at different school levels, by year

	Reading			Mathematics		
	Estimate	Standard error	p value	Estimate	Standard error	p value
Year 1						
Probationary teachers	0.008	0.034	0.823	0.063	0.042	0.138
Nonprobationary teachers	0.008	0.019	0.654	0.047*	0.020	0.021
<i>Difference in effect</i>	-0.001	0.035	0.979	0.016	0.046	0.727
Prior value-added score of 0.2	0.001	0.039	0.982	0.020	0.029	0.498
Prior value-added score of -0.2	0.012	0.040	0.768	0.086*	0.031	0.006
<i>Difference in effect</i>	-0.011	0.069	0.876	-0.066	0.048	0.166
Middle schools	0.061	0.035	0.082	0.064	0.039	0.103
Elementary schools	0.007	0.020	0.742	0.069*	0.022	0.002
<i>Difference in effect</i>	0.054	0.040	0.175	-0.004	0.045	0.922
Year 2						
Probationary teachers	0.051	0.036	0.158	0.081	0.047	0.085
Nonprobationary teachers	0.008	0.028	0.788	0.055	0.035	0.112
<i>Difference in effect</i>	0.044	0.039	0.258	0.026	0.051	0.609
Prior value-added score of 0.2	-0.046	0.056	0.408	0.024	0.048	0.623
Prior value-added score of -0.2	0.094	0.056	0.092	0.085	0.051	0.092
<i>Difference in effect</i>	-0.140	0.100	0.160	-0.062	0.078	0.429
Middle schools	-0.004	0.051	0.940	0.037	0.071	0.605
Elementary schools	0.040	0.029	0.173	0.082*	0.037	0.028
<i>Difference in effect</i>	-0.044	0.058	0.456	-0.046	0.080	0.567

NOTES: Year 1 sample size for reading = 63 schools, 384 teachers, and 13,134 students for the treatment group; 64 schools, 421 teachers, and 15,358 students for the control group. Year 1 sample size for mathematics = 63 schools, 411 teachers, and 13,967 students for the treatment group; 64 schools, 439 teachers, and 15,907 students for the control group. Year 2 sample size for reading = 63 schools, 374 teachers, and 13,962 students for the treatment group; 63 schools, 394 teachers, and 15,423 students for the control group. Year 2 sample size for mathematics = 63 schools, 389 teachers, and 14,186 students for the treatment group; 63 schools, 396 teachers, and 15,809 students for the control group.

The analyses were based on a three-level regression (students within teachers within schools) controlling for random assignment blocks and student background characteristics. The estimates of differential impact represent the differences in impact between probationary and nonprobationary teachers, between teachers whose prior value-added differed by one standard deviation, and between middle school and elementary school teachers, respectively.

None of the estimates are statistically significant at the .05 level (two-tailed).

SOURCES: Spring 2013 and Spring 2014 Teacher Surveys; AIR Value-Added System.

Supporting Exhibits for Analyses of Associations Between Classroom Practice, Principal Leadership, and Student Achievement

Exhibit J.37a. Estimated relationships between classroom practice, principal leadership, and student achievement in reading/ELA, by year

Educator outcome	Year	Standardized coefficient	p value
Classroom practice			
CLASS overall score	Year 2	0.03*	0.002
FFT overall score	Year 2	0.03*	0.003
Principal leadership			
Instructional leadership	Year 1	-0.01	0.615
Instructional leadership	Year 2	0.04	0.088
Teacher-principal trust	Year 1	0.00	0.939
Teacher-principal trust	Year 2	0.02	0.286

NOTES: Year 1 sample size = 63 schools, 384 teachers, and 13,134 students for the treatment group; 64 schools, 421 teachers, and 15,358 students for the control group. Year 2 sample size = 63 schools, 374 teachers, and 13,962 students for the treatment group; 63 schools, 394 teachers, and 15,423 students for the control group.

The analyses were based on a three-level regression (students within teachers within schools) controlling for random assignment blocks and student background characteristics. The standardized coefficient presented in the exhibit represents the change in student achievement score in standard deviation unit that corresponds to a one-standard-deviation increase in the educator outcome.

* Relationship between the education outcome and achievement is statistically significant at the .05 level (two-tailed).

SOURCES: Spring 2014 Classroom Videos; Spring 2013 and Spring 2014 Teacher Surveys; District Administrative Records.

Exhibit J.37b. Estimated relationships between classroom practice, principal leadership, and student achievement in reading/ELA in CLASS districts, by year

Educator outcome	Year	Standardized coefficient	p value
Classroom practice			
CLASS overall score	Year 2	0.02	0.142
FFT overall score	Year 2	0.01	0.363
Principal leadership			
Instructional leadership	Year 1	-0.02	0.251
Instructional leadership	Year 2	0.00	0.942
Teacher-principal trust	Year 1	-0.01	0.501
Teacher-principal trust	Year 2	0.01	0.596

NOTES: Year 1 sample size = 31 schools, 203 teachers, and 7,402 students for the treatment group; 32 schools, 240 teachers, and 8,447 students for the control group. Year 2 sample size = 31 schools, 208 teachers, and 8,059 students for the treatment group; 32 schools, 231 teachers, and 8,997 students for the control group.

The analyses were based on a three-level regression (students within teachers within schools) controlling for random assignment blocks and student background characteristics. The standardized coefficient presented in the exhibit represents the change in student achievement score in standard deviation unit that corresponds to a one-standard-deviation increase in the educator outcome.

None of the relationships between the educator outcome and achievement are statistically significant at the .05 level (two-tailed).

SOURCES: Spring 2014 Classroom Videos; Spring 2013 and Spring 2014 Teacher Surveys; District Administrative Records.

Exhibit J.37c. Estimated relationships between classroom practice, principal leadership, and student achievement in reading/ELA in FFT districts, by year

Educator Outcome	Year	Standardized coefficient	p value
Classroom practice			
CLASS overall score	Year 2	0.03*	0.004
FFT overall score	Year 2	0.04*	0.000
Principal leadership			
Instructional leadership	Year 1	0.01	0.594
Instructional leadership	Year 2	0.02	0.370
Teacher-principal trust	Year 1	0.01	0.593
Teacher-principal trust	Year 2	0.02	0.375

NOTES: Year 1 sample size = 32 schools, 181 teachers, and 5,732 students for the treatment group; 32 schools, 181 teachers, and 6,911 students for the control group. Year 2 sample size = 32 schools, 166 teachers, and 5,903 students for the treatment group; 31 schools, 163 teachers, and 6,426 students for the control group.

The analyses were based on a three-level regression (students within teachers within schools) controlling for random assignment blocks and student background characteristics. The standardized coefficient presented in the exhibit represents the change in student achievement score in standard deviation unit that corresponds to a one-standard-deviation increase in the educator outcome.

* Relationship between the educator outcome and achievement is statistically significant at the .05 level (two-tailed).

SOURCES: Spring 2014 Classroom Videos; Spring 2013 and Spring 2014 Teacher Surveys; District Administrative Records.

Exhibit J.37d. Estimated relationships between classroom practice, principal leadership, and student achievement in mathematics, by year

Educator outcome	Year	Standardized coefficient	p value
Classroom practice			
CLASS overall score	Year 2	0.06*	0.000
FFT overall score	Year 2	0.07*	0.000
Principal leadership			
Instructional leadership	Year 1	-0.02	0.193
Instructional leadership	Year 2	0.01	0.763
Teacher-principal trust	Year 1	-0.04	0.503
Teacher-principal trust	Year 2	0.02	0.372

NOTES: Year 1 sample size = 63 schools, 411 teachers, and 13,967 students for the treatment group; 64 schools, 439 teachers, and 15,907 students for the control group. Year 2 sample size = 63 schools, 389 teachers, and 14,186 students for the treatment group; 63 schools, 396 teachers, and 15,809 students for the control group.

The analyses were based on a three-level regression (students within teachers within schools) controlling for random assignment blocks and student background characteristics. The standardized coefficient presented in the exhibit represents the change in student achievement score in standard deviation unit that corresponds to a one-standard-deviation increase in the educator outcome.

* Relationship between the educator outcome and achievement is statistically significant at the .05 level (two-tailed).

SOURCES: Spring 2014 Classroom Videos; Spring 2013 and Spring 2014 Teacher Surveys; District Administrative Records.

Exhibit J.37e. Estimated relationships between classroom practice, principal leadership, and student achievement in mathematics in CLASS districts, by year

Educator outcome	Year	Standardized coefficient	p value
Classroom practice			
CLASS overall score	Year 2	0.05*	0.004
FFT overall score	Year 2	0.08*	0.000
Principal leadership			
Instructional leadership	Year 1	-0.04	0.075
Instructional leadership	Year 2	-0.01	0.701
Teacher-principal trust	Year 1	-0.02	0.367
Teacher-principal trust	Year 2	-0.00	0.909

NOTES: Year 1 sample size = 31 schools, 232 teachers, and 8,269 students for the treatment group; 32 schools, 257 teachers, and 9,148 students for the control group. Year 2 sample size for mathematics in CLASS districts = 31 schools, 230 teachers, and 8,315 students for the treatment group; 32 schools, 235 teachers, and 8,823 students for the control group.

The analyses were based on a three-level regression (students within teachers within schools) controlling for random assignment blocks and student background characteristics. The standardized coefficient presented in the exhibit represents the change in student achievement score in standard deviation unit that corresponds to a one-standard-deviation increase in the educator outcome.

* Relationship between the educator outcome and achievement is statistically significant at the .05 level (two-tailed).

SOURCES: Spring 2014 Classroom Videos; Spring 2013 and Spring 2014 Teacher Surveys; District Administrative Records.

Exhibit J.37f. Estimated relationships between classroom practice, principal leadership, and student achievement in mathematics in FFT districts, by year

Educator outcome	Year	Standardized coefficient	p value
Classroom practice			
CLASS overall score	Year 2	0.06*	0.000
FFT overall score	Year 2	0.06*	0.000
Principal leadership			
Instructional leadership	Year 1	0.00	0.845
Instructional leadership	Year 2	0.03	0.411
Teacher-principal trust	Year 1	-0.00	0.939
Teacher-principal trust	Year 2	0.05	0.189

NOTES: Year 1 sample size = 32 schools, 179 teachers, and 5,698 students for the treatment group; 32 schools, 182 teachers, and 6,759 students for the control group. Year 2 sample size = 32 schools, 159 teachers, and 5,871 students for the treatment group; 31 schools, 161 teachers, and 6,986 students for the control group.

The analyses were based on a three-level regression (students within teachers within schools) controlling for random assignment blocks and student background characteristics. The standardized coefficient presented in the exhibit represents the change in student achievement score in standard deviation unit that corresponds to a one-standard-deviation increase in the educator outcome.

* Relationship between the educator outcome and achievement is statistically significant at the .05 level (two-tailed).

SOURCES: Spring 2014 Classroom Videos; Spring 2013 and Spring 2014 Teacher Surveys; District Administrative Records.

Appendix K. Sample Reports

Sample CLASS Observation Report

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 

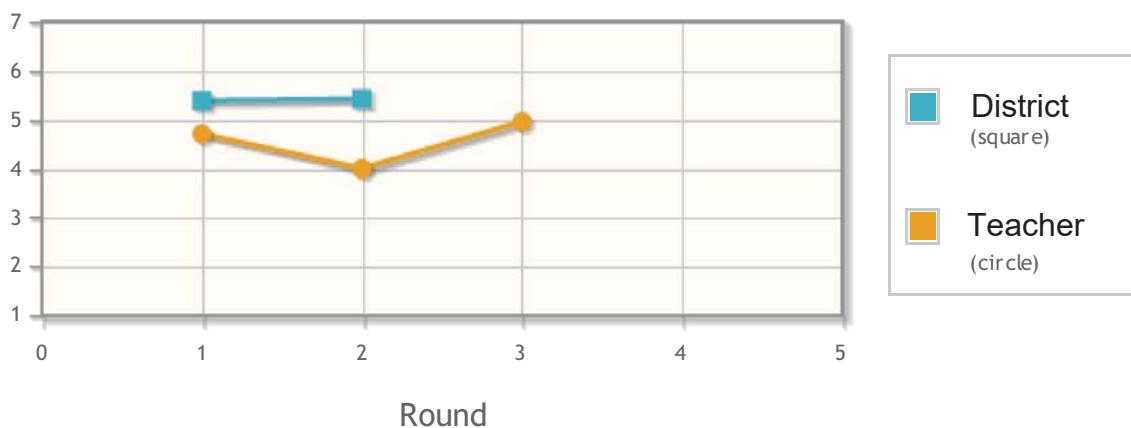
Key:  Ineffective  Developing Effectiveness  Effective  Highly Effective

*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 observation.

average includes only classrooms that received a CLASS

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

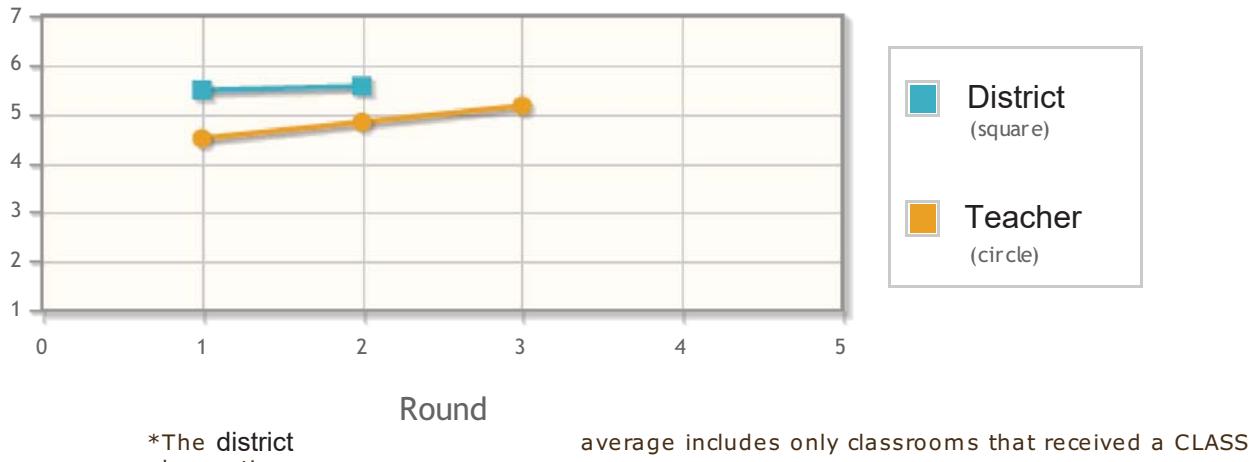
Your overall score was in the **Effective** range. Your areas of strength were indicated in the Emotional Support and Classroom Organization domains as well as Student Engagement. You scored in the highly effective range in these domains however there is always room for continued learning. You demonstrated 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



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.

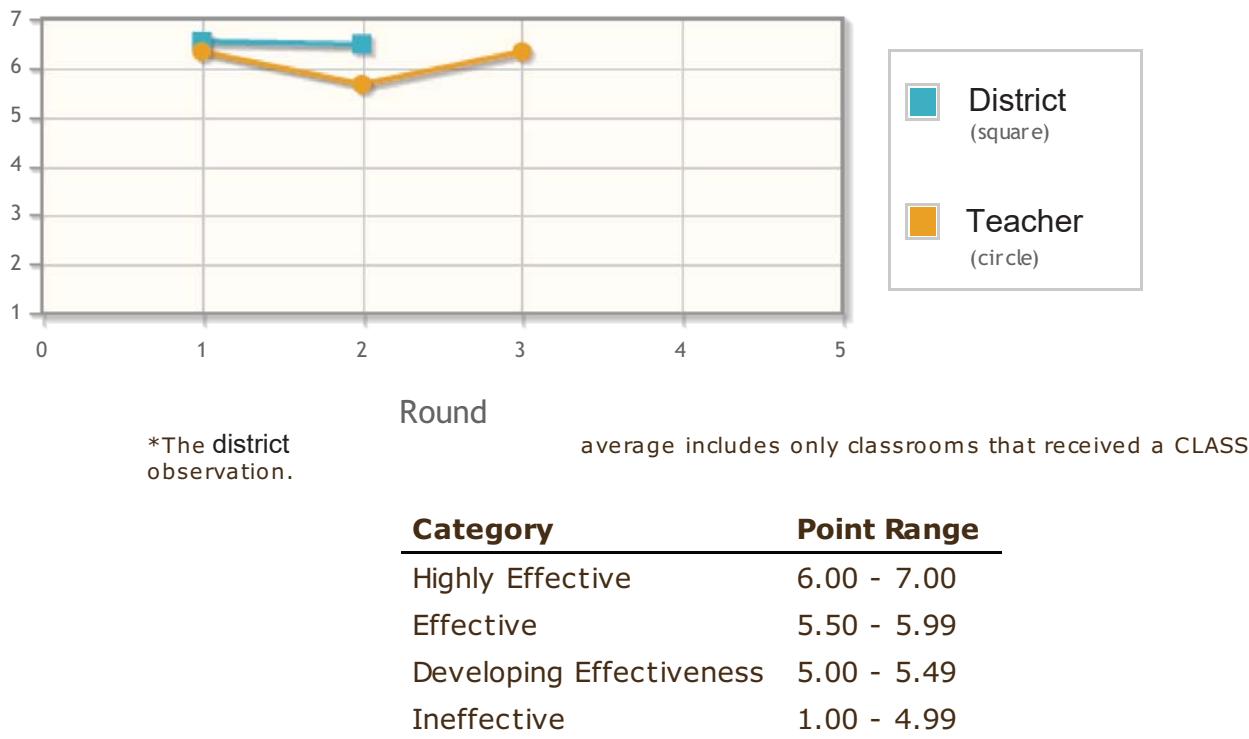
During the observation, the following effective examples were noted:

- 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



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.

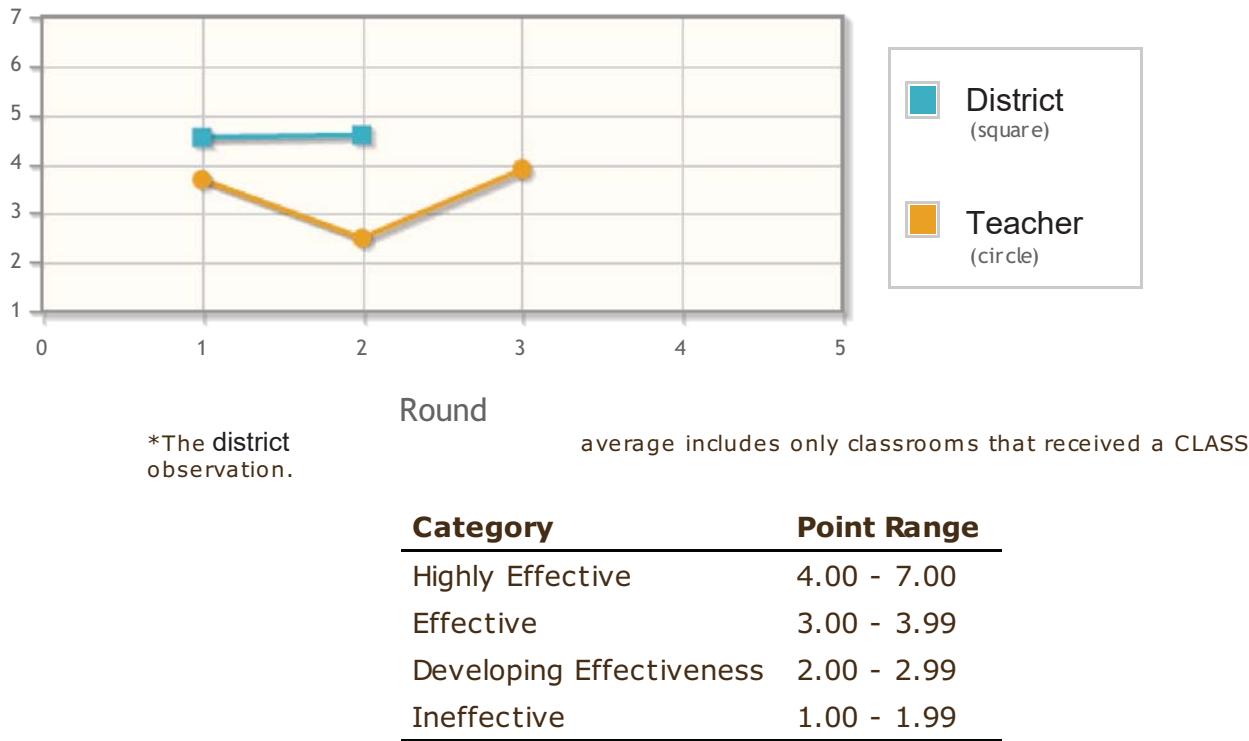
During the observation, the following effective examples were noted:

- 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



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 $6 \times$ any number. If "n" is 1 then $1+ 5 \times 1 = 6 \times 1$. 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.

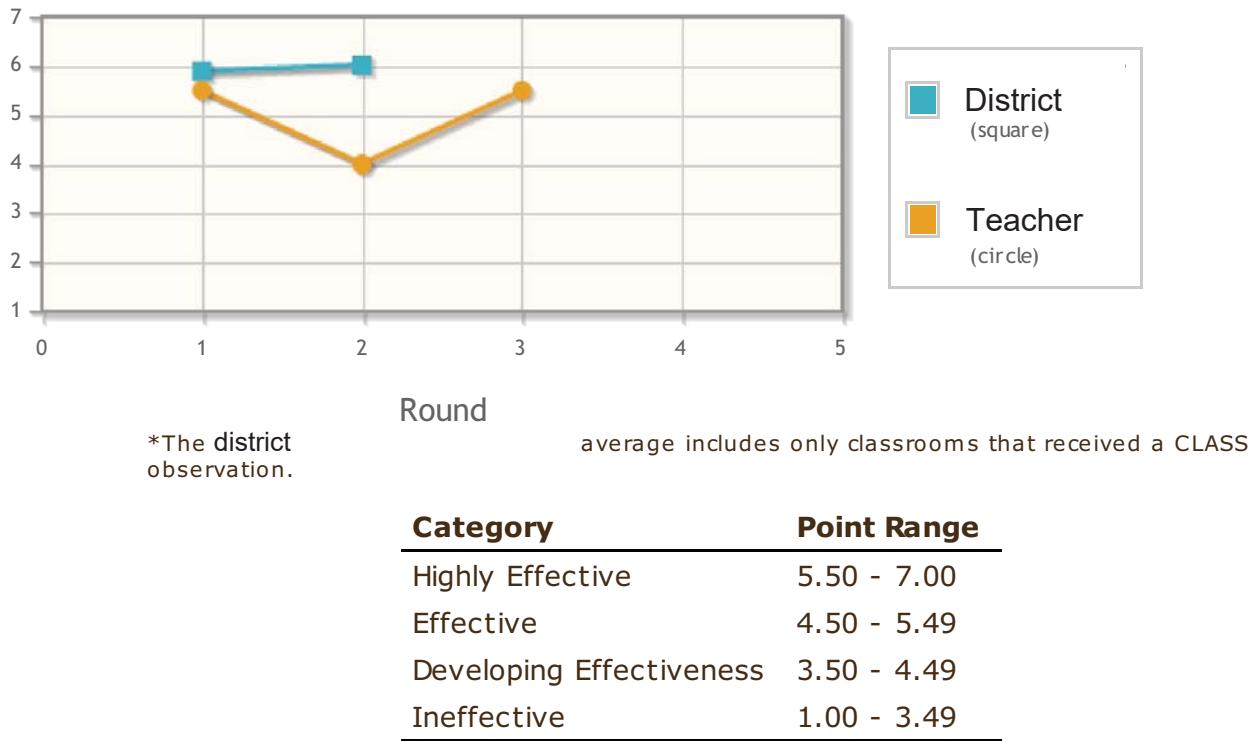
During the observation, the following effective examples were noted:

- 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



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

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

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

2c: Managing classroom procedures

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. (procedures)

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

 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

 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

 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 make 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

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

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

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 comfortable knowing if your students understand the material.

3e: Demonstrating flexibility and responsiveness

Score: NA

Summary

No evidence to score.

Notes

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

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!!

Sample Value-Added Report for Teacher



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	35	35	35
	Mathematics	195	-0.06±0.04	41	41	41	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

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.

Sample Value-Added Report for Principal

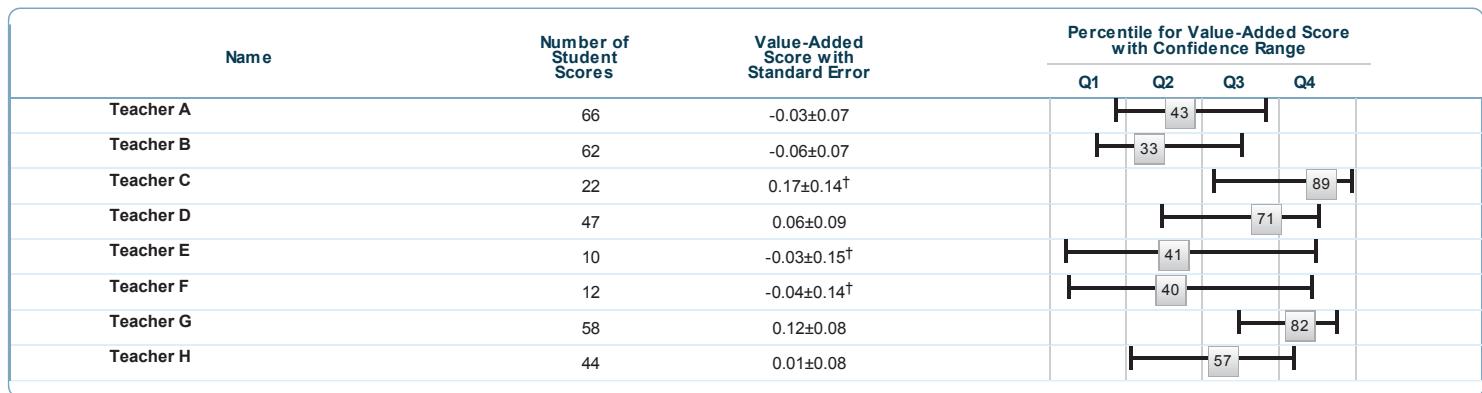


Value-Added Scores for Teachers in School 3

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



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.



Teacher & Leader Evaluation Systems
A Study at American Institutes for Research

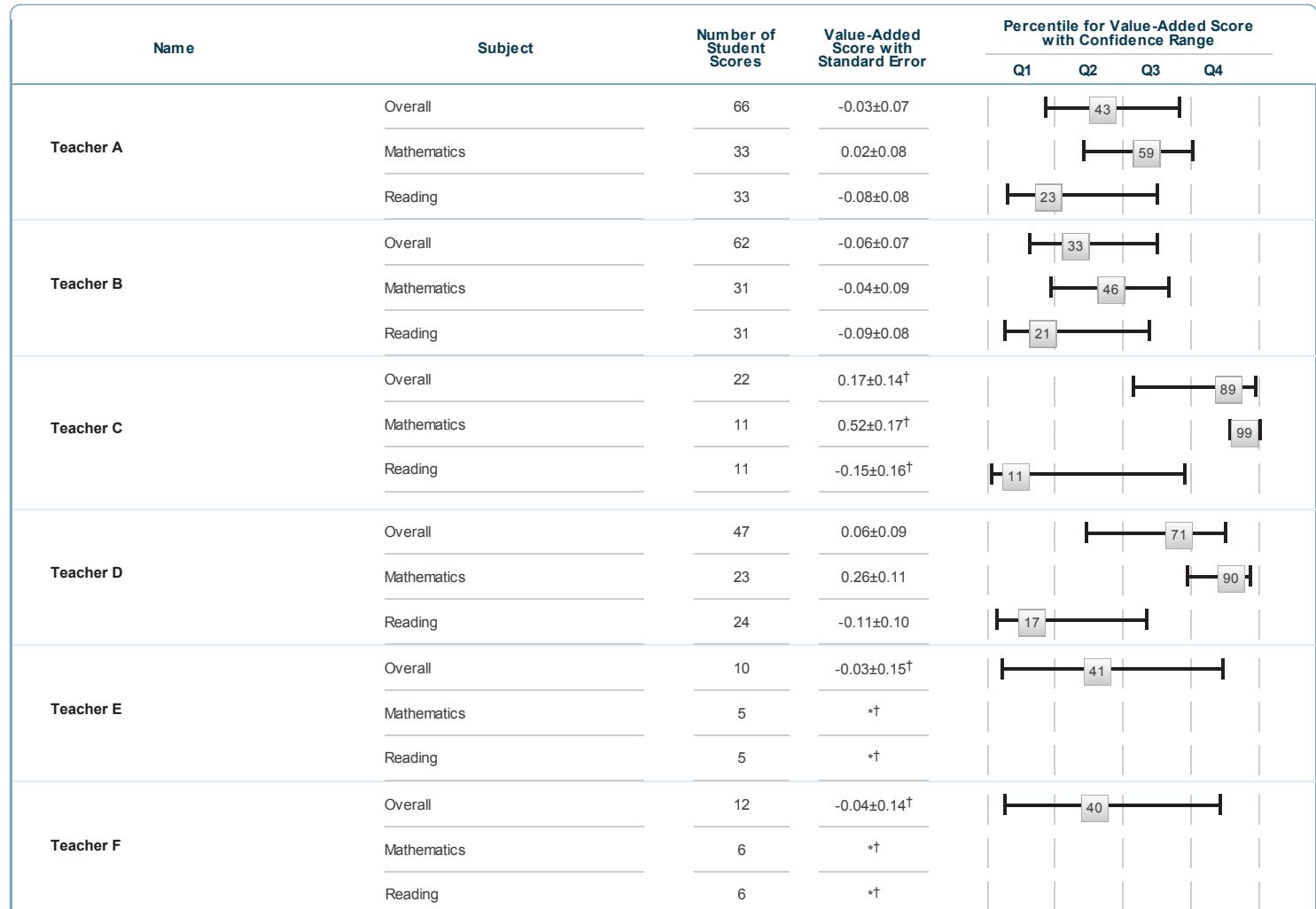


Value-Added Scores for Teachers in School 3 by Subject

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	
	Mathematics		27536	640	0.00±0.01	-0.01±0.11	
	Reading		28393	642	0.00±0.00	0.00±0.10	
School 3	Overall		311	7	0.02±0.03	0.03±0.09	
	Mathematics		157	7	0.08±0.04	0.14±0.12	
	Reading		154	7	-0.04±0.04	-0.06±0.11	



Online Reports

Teacher G	Overall	58	0.12 ± 0.08	
	Mathematics	29	0.17 ± 0.10	
	Reading	29	0.08 ± 0.09	
Teacher H	Overall	44	0.01 ± 0.08	
	Mathematics	24	-0.07 ± 0.10	
	Reading	20	0.08 ± 0.09	

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.



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.

