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 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:


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 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

<table>
<thead>
<tr>
<th>Acknowledgments</th>
<th>vii</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disclosure of Potential Conflicts of Interest</td>
<td>ix</td>
</tr>
<tr>
<td>Executive Summary</td>
<td>ES–1</td>
</tr>
<tr>
<td>Study Overview</td>
<td>ES–2</td>
</tr>
<tr>
<td>Implementation of the Intervention</td>
<td>ES–5</td>
</tr>
<tr>
<td>Contrast in Educators’ Experience of Feedback</td>
<td>ES–9</td>
</tr>
<tr>
<td>Impact on Classroom Practice, Principal Leadership, and Student Achievement</td>
<td>ES–11</td>
</tr>
<tr>
<td>Association Among Classroom Practice, Leadership, and Achievement</td>
<td>ES–16</td>
</tr>
<tr>
<td>Chapter 1. Introduction</td>
<td>1</td>
</tr>
<tr>
<td>Overview of the Intervention</td>
<td>2</td>
</tr>
<tr>
<td>Theory of Action and Research Questions</td>
<td>5</td>
</tr>
<tr>
<td>Overview of Study Design</td>
<td>8</td>
</tr>
<tr>
<td>Chapter 2. Implementation of the Performance Measures and Feedback</td>
<td>21</td>
</tr>
<tr>
<td>The Intervention’s Measures of Teacher Classroom Practice</td>
<td>22</td>
</tr>
<tr>
<td>The Intervention’s Measure of Student Growth</td>
<td>33</td>
</tr>
<tr>
<td>The Intervention’s Measure of Principal Leadership</td>
<td>41</td>
</tr>
<tr>
<td>Summary</td>
<td>47</td>
</tr>
<tr>
<td>Chapter 3. Impact of Performance Feedback on Teacher, Principal, and Student Outcomes</td>
<td>49</td>
</tr>
<tr>
<td>Contrast in Educators’ Experience of Feedback</td>
<td>51</td>
</tr>
<tr>
<td>Impact on Initial Outcomes</td>
<td>55</td>
</tr>
<tr>
<td>Impact on Classroom Practice, Principal Leadership, and Student Achievement</td>
<td>63</td>
</tr>
<tr>
<td>Association Among Classroom Practice, Leadership, and Achievement</td>
<td>77</td>
</tr>
<tr>
<td>Summary</td>
<td>78</td>
</tr>
<tr>
<td>References</td>
<td>79</td>
</tr>
<tr>
<td>Appendix A. Details About the Study Sample</td>
<td>A–1</td>
</tr>
<tr>
<td>Appendix B. Details About Data Collection and Outcome Measures</td>
<td>B–1</td>
</tr>
<tr>
<td>Appendix C. Technical Details About Reliability Estimation</td>
<td>Error! Bookmark not defined.</td>
</tr>
<tr>
<td>Appendix D. Supplemental Findings About the Implementation of the Intervention’s Measures of Classroom Practice</td>
<td>Error! Bookmark not defined.</td>
</tr>
</tbody>
</table>

Appendix F. Supplemental Findings About the Implementation of the Intervention’s Measure of Student Growth .......................................................... F–Error! Bookmark not defined.

Appendix G. Supplemental Findings About the Implementation of the Intervention’s Measure of Principal Leadership .............................................. G–Error! Bookmark not defined.

Appendix H. Technical Details About Analyses Assessing Treatment-Control Differences in Educators’ Experiences and Impacts on Outcomes ....... H–Error! Bookmark not defined.

Appendix I. Supporting Exhibits for Analyses of Educators’ Experiences and Initial Outcomes ....................................................................................... I–Error! Bookmark not defined.


Appendix K. Sample Reports .............................................................. K–Error! Bookmark not defined.
## List of Exhibits

<table>
<thead>
<tr>
<th>Exhibit</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ES.1</td>
<td>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.</td>
</tr>
<tr>
<td>ES.2</td>
<td>Number of feedback instances and duration of oral feedback that principals reported receiving, by treatment status and year.</td>
</tr>
<tr>
<td>ES.3</td>
<td>Average CLASS and FFT scores, based on coding of video-recorded lessons by study team, by treatment status, Year 2.</td>
</tr>
<tr>
<td>ES.4</td>
<td>Average rating of principal instructional leadership and teacher-principal trust, by treatment status and year.</td>
</tr>
<tr>
<td>ES.5</td>
<td>Average reading/English language arts and mathematics achievement, by treatment status and year.</td>
</tr>
<tr>
<td>1.1</td>
<td>Theory of action.</td>
</tr>
<tr>
<td>1.2</td>
<td>District selection and recruitment process.</td>
</tr>
<tr>
<td>1.3</td>
<td>Characteristics of all districts in the United States and districts that participated in the study.</td>
</tr>
<tr>
<td>1.4</td>
<td>Policies and practices for performance feedback to teachers, by district.</td>
</tr>
<tr>
<td>1.5</td>
<td>Policies and practices for performance feedback to principals, by district.</td>
</tr>
<tr>
<td>1.6</td>
<td>Random assignment results, fall 2012.</td>
</tr>
<tr>
<td>2.1</td>
<td>Domains and dimensions of classroom practice for CLASS and FFT.</td>
</tr>
<tr>
<td>2.2</td>
<td>Mean number of feedback sessions treatment teachers received in each study year and in total.</td>
</tr>
<tr>
<td>2.3</td>
<td>Distribution of teachers across CLASS and FFT performance levels for Windows 1 and 4 and for the 4-Window average, Year 2.</td>
</tr>
<tr>
<td>2.4</td>
<td>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.</td>
</tr>
<tr>
<td>2.5</td>
<td>Timeline for estimating value-added scores and delivering student growth reports.</td>
</tr>
<tr>
<td>2.6</td>
<td>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.</td>
</tr>
<tr>
<td>2.7</td>
<td>Percentage of treatment teachers and principals who agreed somewhat or strongly with statements about the student growth reports they viewed, Year 2.</td>
</tr>
<tr>
<td>2.8</td>
<td>VAL-ED core components and key processes.</td>
</tr>
<tr>
<td>2.9</td>
<td>Distribution of treatment principals across performance levels based on VAL-ED overall scores in fall and spring, by year.</td>
</tr>
</tbody>
</table>
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.2. Estimated reliabilities for CLASS overall scores and dimension scores,
Year 2........................................C–Error! Bookmark not defined.

Exhibit C.3. Estimated reliabilities for FFT overall scores and dimension scores, Year 2 .........C–Error! Bookmark not defined.

Exhibit C.4. Estimated variance components and reliabilities for dimension score
differences for CLASS and FFT, Year 2 ..............C–Error! Bookmark not defined.

Exhibit C.5. Estimated reliabilities for value-added scores based on two years of student
growth data, Year 2..........................C–Error! Bookmark not defined.

Exhibit C.6. Estimated variance components and reliability for subject-specific value-
added score differences, Year 2 ............C–Error! Bookmark not defined.

Exhibit C.7. Estimated reliabilities for VAL-ED overall scores and dimension scores, fall
of Year 2 ..............................................C–Error! Bookmark not defined.

Exhibit C.8. Estimated reliabilities for VAL-ED overall scores and dimension scores, spring of Year 2 .........................C–Error! Bookmark not defined.

Exhibit C.9. Estimated variance components and reliabilities for VAL-ED dimension
score differences, fall of Year 2..........C–Error! Bookmark not defined.

Exhibit C.10. Estimated variance components and reliabilities for VAL-ED dimension
score differences, spring of Year 2 ......C–Error! Bookmark not defined.

Exhibit D.1. Percentage of treatment principals who agreed somewhat or strongly with
each statement about the observations they conducted, by year... D–Error! Bookmark not defined.

Exhibit D.2. Mean number of feedback sessions K–3 treatment teachers received, by year
and in total.................................D–Error! Bookmark not defined.

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–Error! Bookmark not defined.

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–Error! Bookmark not defined.

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–Error! Bookmark not defined.

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–Error! Bookmark not defined.

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–Error! Bookmark not defined.
Exhibit D.7a. Distribution of teachers based on their CLASS overall scores in each observation window and the four-window average, by year ....... D–Error! Bookmark not defined.

Exhibit D.7b. Distribution of teachers based on their FFT overall scores in each observation window and the four-window average, by year ....... D–Error! Bookmark not defined.

Exhibit D.7c. Descriptive statistics for CLASS and FFT overall scores in each observation window, by year ................D–Error! Bookmark not defined.

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–Error! Bookmark not defined.

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–Error! Bookmark not defined.

Exhibit D.8. Descriptive statistics for average CLASS and FFT observation scores in each year, by observer type..................D–Error! Bookmark not defined.

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–Error! Bookmark not defined.

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–Error! Bookmark not defined.

Exhibit D.10. Percentage of teachers whose dimension scores spanned one, two, three, or four performance levels, by observation windowD–Error! Bookmark not defined.

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–Error! Bookmark not defined.

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–Error! Bookmark not defined.

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–Error! Bookmark not defined.

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–Error! Bookmark not defined.
Exhibit F.3. Percentage of treatment teachers who agreed somewhat or strongly with statements about the student growth reports they viewed, Year 2.

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

Exhibit G.1. Definitions of VAL-ED core components and key processes...

Exhibit G.2. Sample VAL-ED survey items...

Exhibit G.3. Results overview from a sample VAL-ED report...

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

Exhibit G.5. Summary of component-by-process scores from a sample VAL-ED report...

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

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

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

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...

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.

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.

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.

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.
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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

Exhibit I.12b. Teachers’ prior value-added percentile for teachers with self-appraisals in different categories, by subject, Year 1.
Exhibit I.12c. Teachers’ prior value-added percentile for teachers with self-appraisals in different categories, by subject, Year 2.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.
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. 

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.

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.

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.

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.

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.

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.

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.

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.

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.

Exhibit J.6. Background characteristics of students in Year 1 reading/ELA achievement impact sample, by treatment status.

Exhibit J.7. Background characteristics of students in Year 1 reading/ELA achievement impact sample in CLASS and FFT districts, by treatment status.
Exhibit J.8. Background characteristics of students in Year 1 mathematics achievement impact sample, by treatment status .......... J–Error! Bookmark not defined.

Exhibit J.9. Background characteristics of students in Year 1 mathematics achievement impact sample in CLASS and FFT districts, by treatment status .. J–Error! Bookmark not defined.


Exhibit J.11. Background characteristics of students in Year 2 reading/ELA achievement impact sample in CLASS and FFT districts, by treatment status .. J–Error! Bookmark not defined.


Exhibit J.13. Background characteristics of students in Year 2 mathematics achievement impact sample in CLASS and FFT districts, by treatment status .. J–Error! Bookmark not defined.

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–Error! Bookmark not defined.

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–Error! Bookmark not defined.

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–Error! Bookmark not defined.

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–Error! Bookmark not defined.

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–Error! Bookmark not defined.

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–Error! Bookmark not defined.


Exhibit J.22. Average rating of principal instructional leadership, by treatment status, district, and year .............................................. J–Error! Bookmark not defined.


Exhibit J.27. Average reading/ELA and mathematics achievement in CLASS and FFT districts, by treatment status and year .... J–Error! Bookmark not defined.


Exhibit J.30. Average reading/ELA and mathematics achievement without covariate adjustment, by treatment status and year J–Error! Bookmark not defined.

Exhibit J.31. Average reading/ELA and mathematics achievement without covariate adjustment in CLASS and FFT districts, by treatment status and year ............................................. J–Error! Bookmark not defined.

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–Error! Bookmark not defined.

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–Error! Bookmark not defined.

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–Error! Bookmark not defined.


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–Error! Bookmark not defined.
Exhibit J.37a. Estimated relationships between classroom practice, principal leadership, and student achievement in reading/ELA, by year J–Error! Bookmark not defined.

Exhibit J.37b. Estimated relationships between classroom practice, principal leadership, and student achievement in reading/ELA in CLASS districts, by year J–Error! Bookmark not defined.

Exhibit J.37c. Estimated relationships between classroom practice, principal leadership, and student achievement in reading/ELA in FFT districts, by year J–Error! Bookmark not defined.

Exhibit J.37d. Estimated relationships between classroom practice, principal leadership, and student achievement in mathematics, by year J–Error! Bookmark not defined.

Exhibit J.37e. Estimated relationships between classroom practice, principal leadership, and student achievement in mathematics in CLASS districts, by year J–Error! Bookmark not defined.

Exhibit J.37f. Estimated relationships between classroom practice, principal leadership, and student achievement in mathematics in FFT districts, by year J–Error! Bookmark not defined.
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. 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. 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. 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. 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. 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. Among teachers with value-added scores in both reading/English language arts 12

---

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 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.

Executive Summary ES-7 Year 2 Report
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. 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. 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. All but two principals in Year 2 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. 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.
**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

<table>
<thead>
<tr>
<th>Year 1</th>
<th>Year 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Reading</strong></td>
<td><strong>Mathematics</strong></td>
</tr>
<tr>
<td>Treatment</td>
<td>51.8*</td>
</tr>
<tr>
<td>Control</td>
<td>49.9</td>
</tr>
</tbody>
</table>

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. 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. For example, recent research suggests that providing more frequent, specific feedback on classroom practice may lead to improvements in teacher performance and student achievement.

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. 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. 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

---

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).

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. 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).

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. The use of two different observation systems was intended to make the study findings more robust.

---

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.

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. 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

**Measurement of performance and delivery of feedback**
*research question 1*
- Measurement of performance that is frequent and systematic
- Delivery of feedback on performance that is frequent and systematic

**Experiences of feedback**
*research question 3*
- Number of feedback sessions with ratings and written narrative
- Duration of feedback sessions

**Initial outcomes**
- Discussion of the measured dimensions
- Educators’ perceptions of their own performance
- Interest in improving on the measured dimensions
- Participation in professional development on the measured dimensions

**Mechanisms**
- increased knowledge, skills, and effort among educators who remain in their positions during the intervention

**Educator practices**
*research question 4*
- Teacher classroom practice
  - CLASS
  - FFT
- Principal leadership
  - Instructional leadership
  - Teacher-principal trust

**Student achievement**
*research question 5*
- Reading/English language arts
- Mathematics
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. 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).
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

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

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

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. 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

<table>
<thead>
<tr>
<th>District ID and assigned classroom observation system for intervention</th>
<th>Districts’ feedback on teacher classroom practice</th>
<th>Districts’ feedback on student growth</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Frequency of observation with feedback&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Use of staff not based at the school as observers</td>
</tr>
<tr>
<td></td>
<td>Probationary teachers&lt;sup&gt;b&lt;/sup&gt;</td>
<td>Nonprobationary teachers&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>1</td>
<td>CLASS</td>
<td>1 per year</td>
</tr>
<tr>
<td>2</td>
<td>CLASS</td>
<td>1 per year</td>
</tr>
<tr>
<td>3</td>
<td>CLASS</td>
<td>1 per year</td>
</tr>
<tr>
<td>4</td>
<td>CLASS</td>
<td>3 per year</td>
</tr>
<tr>
<td>5</td>
<td>FFT</td>
<td>2 per year</td>
</tr>
<tr>
<td>6</td>
<td>FFT</td>
<td>2 per year</td>
</tr>
<tr>
<td>7</td>
<td>FFT</td>
<td>2 per year</td>
</tr>
<tr>
<td>8</td>
<td>FFT</td>
<td>1 per year</td>
</tr>
<tr>
<td><strong>Overall average</strong></td>
<td>1.6 per year</td>
<td>0.5 per year</td>
</tr>
</tbody>
</table>

**NOTES:**

<sup>a</sup>Number of observations shown is the minimum required under each district’s evaluation system. Administrators could observe more frequently at their discretion.

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

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

**SOURCE:** District 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.

<table>
<thead>
<tr>
<th>District ID and assigned classroom observation system for intervention</th>
<th></th>
<th>Districts’ feedback on principal leadership</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>Use of teacher survey as input in principal evaluation</td>
<td>Rating instrument with multiple dimensions</td>
<td>Performance on each dimension rated using three or more performance levels</td>
</tr>
<tr>
<td>1 CLASS</td>
<td>1 per year</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>2 CLASS</td>
<td>1 per year</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>3 CLASS</td>
<td>1 per year</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>4 CLASS</td>
<td>1 per year</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>5 FFT</td>
<td>1 per year</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>6 FFT</td>
<td>1 per year</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>7 FFT</td>
<td>1 per year</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>8 FFT</td>
<td>1 per year</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

**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.)

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

---

<table>
<thead>
<tr>
<th>Treatment status</th>
<th>Number of schools</th>
<th>Number of teachers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
<td>Elementary schools</td>
</tr>
<tr>
<td>Treatment</td>
<td>63</td>
<td>49</td>
</tr>
<tr>
<td>Control</td>
<td>64</td>
<td>48</td>
</tr>
<tr>
<td>Total</td>
<td>127</td>
<td>97</td>
</tr>
</tbody>
</table>

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.\footnote{We video-recorded two lessons for some teachers and one for others to achieve the desired precision while minimizing cost and burden.} We coded each of the videos using the CLASS
and FFT. \footnote{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.} 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). \footnote{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.}

**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.
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. 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.

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.\textsuperscript{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.\textsuperscript{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.\textsuperscript{47,48}

\textsuperscript{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.

\textsuperscript{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.

\textsuperscript{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.

\textsuperscript{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.) 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

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

**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. 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. 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. (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% 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.  

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

<table>
<thead>
<tr>
<th></th>
<th>Year 1 feedback sessions (Year 1 Impact Sample)</th>
<th>Year 2 feedback sessions (Year 2 Impact Sample)</th>
<th>Year 1 and 2 feedback sessions (Year 2 Impact Sample)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CLASS</td>
<td>FFT</td>
<td>All</td>
</tr>
<tr>
<td>Mean number of feedback sessions received</td>
<td>3.5</td>
<td>3.9</td>
<td>3.7</td>
</tr>
</tbody>
</table>

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. 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; 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.

<table>
<thead>
<tr>
<th></th>
<th>CLASS Instrument</th>
<th>FFT Instrument</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Highly Effective</td>
<td>Score of 3.50</td>
</tr>
<tr>
<td>Window 1</td>
<td>(5.00 to 7.00)</td>
<td>to 4.00</td>
</tr>
<tr>
<td></td>
<td>Effective (3.50</td>
<td>Score of 2.50</td>
</tr>
<tr>
<td>Window 4</td>
<td>to 4.99)</td>
<td>to 3.49</td>
</tr>
<tr>
<td></td>
<td>Developing Effectiveness (2.50 to 3.49)</td>
<td>Score of 1.50 to 2.49</td>
</tr>
<tr>
<td></td>
<td>Ineffective^a (1.00 to 2.49)</td>
<td>Score of 1.00 to 1.49</td>
</tr>
<tr>
<td>4-Window Average</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>70%</td>
<td>10%</td>
</tr>
<tr>
<td></td>
<td>90%</td>
<td>19%</td>
</tr>
<tr>
<td></td>
<td>85%</td>
<td>86%</td>
</tr>
<tr>
<td></td>
<td>1%</td>
<td>5%</td>
</tr>
<tr>
<td></td>
<td>28%</td>
<td>78%</td>
</tr>
<tr>
<td></td>
<td>9%</td>
<td>77%</td>
</tr>
<tr>
<td></td>
<td>15%</td>
<td>8%</td>
</tr>
<tr>
<td></td>
<td>1%</td>
<td>6%</td>
</tr>
<tr>
<td>Window 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Window 4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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.\textsuperscript{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.\textsuperscript{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.\textsuperscript{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.\textsuperscript{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

\textsuperscript{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.

\textsuperscript{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.

\textsuperscript{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.

\textsuperscript{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.

\textsuperscript{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 understate 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. 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. 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.
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.)
Chapter 2. Implementation of the Performance Measures and Feedback

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

**Teachers: Compared to previous feedback from the district, the feedback I received from the study's CLASS/FFT observations ...**

- **Was harder to understand**: 36%
- **Was more objective**: 73%
- **Was more specific about what constitutes high-quality teaching**: 79%
- **Was more focused on specific things I did during the observation**: 84%
- **Provided me with clearer ideas about how my teaching could improve**: 73%
- **Had more detailed written narratives**: 78%
- **Had more useful feedback sessions**: 65%
- **Was more critical of my performance**: 65%

**Principals: The CLASS/FFT rating system ...**

- **Does a good job distinguishing effective from ineffective teaching**: ≥ 95%
- **Is fair to all teachers, regardless of their personal characteristics**: 92%
- **Is fair to all teachers, regardless of the characteristics of the students they teach**: 94%
- **Does NOT accurately reflect the quality of an individual's teaching**: 24%

**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.

<table>
<thead>
<tr>
<th>Exhibit 2.5. Timeline for estimating value-added scores and delivering student growth reports</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wave 1</td>
</tr>
<tr>
<td>Value-added scores estimated for these years</td>
</tr>
<tr>
<td>Wave 2</td>
</tr>
<tr>
<td>Wave 3</td>
</tr>
</tbody>
</table>

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.\textsuperscript{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.\textsuperscript{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**

\textsuperscript{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.

\textsuperscript{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).

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

\textsuperscript{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). 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.) 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.

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).\textsuperscript{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.\textsuperscript{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.

\textsuperscript{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.

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

\textsuperscript{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.
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).\textsuperscript{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.\textsuperscript{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.)

\textsuperscript{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.

\textsuperscript{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.)

\textsuperscript{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).
Chapter 2. Implementation of the Performance Measures and Feedback

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**
- The VA report was easy to understand: 52%
- I understand what I would need to do to improve my VA score: 55%
- The VA score is a good measure of how well students learned what I taught last year: 48%
- The VA score is a good measure of my overall performance as a teacher: 43%
- The information as an indicator of teacher effectiveness is fair to all teachers, regardless of the personal characteristics of the students they teach: 42%
- The information as an indicator of teacher effectiveness is fair to all teachers, regardless of the prior achievement of the students they teach: 41%

**Principals**
- The information as an indicator of teacher effectiveness is fair to all teachers, regardless of the personal characteristics of the students they teach: 75%
- The information as an indicator of teacher effectiveness is fair to all teachers, regardless of the prior achievement of students they teach: 72%
- The VA score is a good measure of how well students learned what teachers in my school taught during the year: 74%
- The VA report does a good job distinguishing effective from ineffective teachers: 70%
- The VA report accurately reflects the quality of teachers who taught in this school: 72%

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.

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.”

<table>
<thead>
<tr>
<th>Exhibit 2.8. VAL-ED core components and key processes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Core components</strong></td>
</tr>
<tr>
<td>• High standards for student learning</td>
</tr>
<tr>
<td>• Rigorous curriculum</td>
</tr>
<tr>
<td>• Quality instruction</td>
</tr>
<tr>
<td>• Culture of learning and professional behavior</td>
</tr>
<tr>
<td>• Connections to external communities</td>
</tr>
<tr>
<td>• Systemic performance accountability</td>
</tr>
</tbody>
</table>

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. (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). (See appendix exhibit G.3 for a screen shot from a sample VAL-ED report that includes the performance-level...

---

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).
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. Two principals in Year 2 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**

<table>
<thead>
<tr>
<th>Year</th>
<th>Fall</th>
<th>Spring</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year 1</td>
<td>8%</td>
<td>22%</td>
</tr>
<tr>
<td></td>
<td>43%</td>
<td>37%</td>
</tr>
<tr>
<td></td>
<td>27%</td>
<td>14%</td>
</tr>
<tr>
<td>Year 2</td>
<td>11%</td>
<td>46%</td>
</tr>
<tr>
<td></td>
<td>41%</td>
<td>30%</td>
</tr>
<tr>
<td></td>
<td>17%</td>
<td>11%</td>
</tr>
</tbody>
</table>

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).


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. 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. 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. 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.) The improved reliability in spring reflects greater average agreement among the respondent groups.

---

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.)
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

<table>
<thead>
<tr>
<th>Correlation</th>
<th>Fall Year 1</th>
<th>Spring Year 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Principal and supervisor</td>
<td>.08</td>
<td>.27*</td>
</tr>
<tr>
<td>Principal and teachers</td>
<td>.06</td>
<td>.26*</td>
</tr>
<tr>
<td>Supervisor and teachers</td>
<td>.27*</td>
<td>.38*</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Correlation</th>
<th>Fall Year 2</th>
<th>Spring Year 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Principal and supervisor</td>
<td>.01</td>
<td>.23</td>
</tr>
<tr>
<td>Principal and teachers</td>
<td>.34*</td>
<td>.28*</td>
</tr>
<tr>
<td>Supervisor and teachers</td>
<td>.08</td>
<td>.38*</td>
</tr>
</tbody>
</table>

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 $p < .05$.


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.\(^9\) 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.

---

\(^9\) 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

Compared to previous feedback from the district, the feedback I received from the VAL-ED observations ...

- **Was harder to understand**: 34%
- **Was more objective**: 73%
- **Was more specific about what constitutes high-quality leadership**: 78%
- **Provided me with clearer ideas about how my leadership could improve**: 75%
- **Was less comprehensive, ignoring some aspects of my role as principal**: 55%
- **Was more critical of my performance**: 58%

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 than 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.\textsuperscript{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.\textsuperscript{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

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

\textsuperscript{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.

\textsuperscript{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. (See appendix exhibits 1.1a and 1b.) The results were particularly pronounced for nonprobationary teachers. 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.) 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. 

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 I.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.)\textsuperscript{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).\textsuperscript{102}

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

![Exhibit 3.3](image)

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

\textsuperscript{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.

\textsuperscript{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 areas are 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).

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. 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**

<table>
<thead>
<tr>
<th>Areas of practice aligned to CLASS/FFT</th>
<th>Treatment</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discussed with supervisor</td>
<td>89%</td>
<td>78%</td>
</tr>
<tr>
<td>Interested in improving</td>
<td>57%</td>
<td>56%</td>
</tr>
<tr>
<td>Received professional development</td>
<td>63%</td>
<td>68%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Areas of practice not aligned to CLASS/FFT</th>
<th>Treatment</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discussed with supervisor</td>
<td>76%</td>
<td>80%</td>
</tr>
<tr>
<td>Interested in improving</td>
<td>55%</td>
<td>60%</td>
</tr>
<tr>
<td>Received professional development</td>
<td>79%</td>
<td>84%</td>
</tr>
</tbody>
</table>

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.

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. 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).

---

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

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 exhibit 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.)

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, 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.

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

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. 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.

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.

**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-
point CLASS scale, compared with 4.39 for control teachers.\textsuperscript{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.\textsuperscript{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.

\footnotesize{\textsuperscript{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.)

\textsuperscript{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: The average CLASS overall score was 4.50 for treatment teachers, compared with 4.39 for control teachers. 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).

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.

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. 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). 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

**CLASS Districts**

<table>
<thead>
<tr>
<th>Year</th>
<th>Instructional Leadership</th>
<th>Teacher-Principal Trust</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Treatment 3.41</td>
<td>Control 3.30</td>
</tr>
<tr>
<td></td>
<td>Treatment 3.23</td>
<td>Control 3.05</td>
</tr>
<tr>
<td>2</td>
<td>Treatment 3.36</td>
<td>Control 3.33</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**FFT Districts**

<table>
<thead>
<tr>
<th>Year</th>
<th>Instructional Leadership</th>
<th>Teacher-Principal Trust</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Treatment 3.13</td>
<td>Control 3.09</td>
</tr>
<tr>
<td></td>
<td>Treatment 3.14*</td>
<td>Control 2.90</td>
</tr>
<tr>
<td>2</td>
<td>Treatment 3.33*</td>
<td>Control 3.09</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Control 3.26*</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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. \((\text{See exhibit 3.12.})\) 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. \((\text{See appendix exhibits A.11 and 12 for a description of student entries and exits.})\) The treatment and control students included in the achievement impact analyses did not differ in demographic characteristics or prior achievement. \((\text{See appendix exhibits J.6–13.})\) 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. \((\text{See appendix exhibits J.32 and 33.})\)

\(^{124}\) 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. \((\text{See appendix exhibit J.36.})\)

\(^{125}\) 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.
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. 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. (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 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


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

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

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).

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

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

**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

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

**Number of schools**

<table>
<thead>
<tr>
<th>CLASS districts</th>
<th>FFT districts</th>
</tr>
</thead>
<tbody>
<tr>
<td>63</td>
<td>64</td>
</tr>
</tbody>
</table>

**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

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

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

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

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

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

**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).


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

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

**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

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

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

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

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

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Treatment group</th>
<th>Control group</th>
<th>Estimated difference</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Years of experience in district</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean number of years</td>
<td>9.6</td>
<td>10.3</td>
<td>-0.7</td>
<td>.252</td>
</tr>
<tr>
<td>Three years or fewer (percentage)</td>
<td>25.8</td>
<td>24.8</td>
<td>1.0</td>
<td>.752</td>
</tr>
<tr>
<td>Four to 10 years (percentage)</td>
<td>37.9</td>
<td>34.8</td>
<td>3.0</td>
<td>.357</td>
</tr>
<tr>
<td>Eleven to 20 years (percentage)</td>
<td>23.9</td>
<td>25.4</td>
<td>-1.4</td>
<td>.597</td>
</tr>
<tr>
<td>More than 20 years (percentage)</td>
<td>12.3</td>
<td>14.8</td>
<td>-2.5</td>
<td>.308</td>
</tr>
<tr>
<td>Master’s degree or higher (percentage)</td>
<td>43.9</td>
<td>46.1</td>
<td>-2.1</td>
<td>.396</td>
</tr>
</tbody>
</table>

**Number of grade 4–8 teachers**

<table>
<thead>
<tr>
<th></th>
<th>Treatment group</th>
<th>Control group</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>575</td>
<td>594</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**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

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Treatment group</th>
<th>Control group</th>
<th>Estimated difference</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Years of experience in district</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean number of years</td>
<td>10.5</td>
<td>9.8</td>
<td>0.6</td>
<td>.255</td>
</tr>
<tr>
<td>Three years or fewer (percentage)</td>
<td>20.1</td>
<td>23.4</td>
<td>-3.4</td>
<td>.186</td>
</tr>
<tr>
<td>Four to 10 years (percentage)</td>
<td>41.3</td>
<td>38.1</td>
<td>3.1</td>
<td>.224</td>
</tr>
<tr>
<td>Eleven to 20 years (percentage)</td>
<td>23.3</td>
<td>24.6</td>
<td>-1.4</td>
<td>.545</td>
</tr>
<tr>
<td>More than 20 years (percentage)</td>
<td>15.4</td>
<td>14.1</td>
<td>1.3</td>
<td>.511</td>
</tr>
<tr>
<td>Master’s degree or higher (percentage)</td>
<td>31.7</td>
<td>33.8</td>
<td>-2.1</td>
<td>.283</td>
</tr>
</tbody>
</table>

**Number of grade 4–8 teachers**

<table>
<thead>
<tr>
<th></th>
<th>Treatment group</th>
<th>Control group</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>337</td>
<td>344</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**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

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

**Number of grade 4–8 teachers**

|                          | Treatment group | Control group | 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

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Treatment group</th>
<th>Control group</th>
<th>Estimated difference</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Students eligible for free or reduced-price lunch (percentage)</td>
<td>60.2</td>
<td>61.6</td>
<td>-1.4</td>
<td>.351</td>
</tr>
<tr>
<td>Race/ethnicity (percentage)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>44.2</td>
<td>43.1</td>
<td>1.1</td>
<td>.334</td>
</tr>
<tr>
<td>Black or African American</td>
<td>3.1</td>
<td>3.4</td>
<td>-0.3</td>
<td>.439</td>
</tr>
<tr>
<td>Hispanic</td>
<td>47.8</td>
<td>48.3</td>
<td>-0.6</td>
<td>.647</td>
</tr>
<tr>
<td>Asian/Pacific Islander</td>
<td>2.5</td>
<td>2.5</td>
<td>0.0</td>
<td>.991</td>
</tr>
<tr>
<td>Other</td>
<td>2.5</td>
<td>2.9</td>
<td>-0.4</td>
<td>.651</td>
</tr>
<tr>
<td>Female (percentage)</td>
<td>49.1</td>
<td>48.3</td>
<td>0.8</td>
<td>.204</td>
</tr>
<tr>
<td>English language learners (percentage)</td>
<td>15.6</td>
<td>16.9</td>
<td>-1.3</td>
<td>.360</td>
</tr>
<tr>
<td>Students with disabilities (percentage)</td>
<td>11.7</td>
<td>9.8</td>
<td>1.8</td>
<td>.159</td>
</tr>
<tr>
<td>Student achievement on state assessment (standardized)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011–12 Grade 4–8 reading achievement</td>
<td>-0.029</td>
<td>0.022</td>
<td>-0.051</td>
<td>.111</td>
</tr>
<tr>
<td>2011–12 Grade 4–8 mathematics achievement</td>
<td>-0.009</td>
<td>-0.006</td>
<td>-0.003</td>
<td>.932</td>
</tr>
</tbody>
</table>

**Number of grade 4–8 students**

|                          | Treatment group | Control group | 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

Spring 2013
Principal Impact Analysis Sample
(total number of principals in spring 2013)
N = 127

Spring 2013
Principals in treatment schools
N = 63
Left treatment
N = 13
Incoming principals
N = 10
Moved to treatment from control
N = 3

Spring 2014
Principals in treatment schools
N = 63
10+3 = 13 principals (21%)
not present in spring 2013

Spring 2014
Principals in control schools
N = 64
Left control
N = 10
Incoming principals
N = 9
Moved to control from treatment
N = 0

Spring 2014
Principal Impact Analysis Sample
(total number of principals in spring 2014)
N = 126
13+9 = 22 principals (17%)
not present in spring 2013

SOURCE: Study Records.
Exhibit A.6. Teacher turnover across study years

**Spring 2013**
Teacher Impact Analysis Sample
(total number of teachers in spring 2013)
\[ N = 1,091 \]

- **Spring 2013**
  Teacher in treatment schools
  \[ N = 527 \]
  - Left treatment
  \[ N = 134 \]
  - Incoming teachers
  \[ N = 102 \]
  - Moved to treatment from control
  \[ N = 9 \]

- **Spring 2013**
  Teachers in control schools
  \[ N = 564 \]
  - Left control
  \[ N = 181 \]
  - Incoming teachers
  \[ N = 143 \]
  - Moved to control from treatment
  \[ N = 8 \]

**Spring 2014**
Teachers in treatment schools
\[ N = 504 \]
102 + 9 = 111 teachers (22%)
not present in spring 2013

**Spring 2014**
Teachers in control schools
\[ N = 534 \]
143 + 8 = 151 teachers (28%)
not present in spring 2013

**Spring 2014**
Teacher Impact Analysis Sample
(total number of teachers in spring 2014)
\[ N = 1,038 \]
111 + 151 = 262 Teachers (25%)
not present in spring 2013

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

Spring 2013
Student Impact Analysis Sample
(total number of students in spring 2013)
N = 28,492

Spring 2013
Students in treatment schools
N = 13,134

Left treatment
N = 6,041

Incoming students
N = 5,892

Moved to treatment from control
N = 977

Spring 2014
Students in treatment schools
N = 13,962
5,892+977 = 6,869 students (49%) not present in spring 2013

Spring 2014
Students in control schools
N = 15,423
5,997+1,115 = 7,112 students (46%) not present in spring 2013

Spring 2014
Student Impact Analysis Sample
(total number of students in spring 2014)
N = 29,385
6,869+7,112 = 13,981 Students (48%) not present in spring 2013

Left control
N = 7,047

Incoming students
N = 5,997

Moved to control from treatment
N = 1,115

SOURCE: Study Records.
Exhibit A.8. Student turnover across study years, mathematics achievement impact sample

Spring 2013
Student Impact Analysis Sample
(total number of students in spring 2013)
N = 29,874

Spring 2013
Students in treatment schools
N = 13,967

Left treatment
N = 6,434

Incoming students
N = 5,668

Moved to treatment from control
N = 985

Spring 2014
Students in treatment schools
N = 14,186
5,668+985 = 6,653 students (47%)
not present in spring 2013

Spring 2014
Students in control schools
N = 15,809
5,697+1,262 = 6,959 students (44%)
not present in spring 2013

Spring 2013
Students in control schools
N = 15,907

Left control
N = 7,057

Incoming students
N = 5,697

Moved to control from treatment
N = 1,262

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

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

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.

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Year</th>
<th>Realized MDES</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Classroom practice</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CLASS overall score</td>
<td>Year 2</td>
<td>0.18</td>
</tr>
<tr>
<td>FFT overall score</td>
<td>Year 2</td>
<td>0.14</td>
</tr>
<tr>
<td><strong>Principal leadership</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Instructional leadership</td>
<td>Year 1</td>
<td>0.29</td>
</tr>
<tr>
<td>Teacher-principal trust</td>
<td>Year 1</td>
<td>0.26</td>
</tr>
<tr>
<td>Instructional leadership</td>
<td>Year 2</td>
<td>0.26</td>
</tr>
<tr>
<td>Teacher-principal trust</td>
<td>Year 2</td>
<td>0.28</td>
</tr>
<tr>
<td><strong>Student achievement</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reading</td>
<td>Year 1</td>
<td>0.05</td>
</tr>
<tr>
<td>Mathematics</td>
<td>Year 1</td>
<td>0.05</td>
</tr>
<tr>
<td>Reading</td>
<td>Year 2</td>
<td>0.06</td>
</tr>
<tr>
<td>Mathematics</td>
<td>Year 2</td>
<td>0.09</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Observer information sheets and certification results</td>
<td>Summer</td>
<td>Summer</td>
<td>End of year</td>
<td></td>
</tr>
<tr>
<td>Study-hired observer questionnaire</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CLASS/FFT online system records</td>
<td>Throughout school year</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VAL-ED online system records</td>
<td>November</td>
<td></td>
<td>April</td>
<td></td>
</tr>
<tr>
<td>AIR online system records</td>
<td></td>
<td></td>
<td>End of year</td>
<td></td>
</tr>
<tr>
<td>District interviews</td>
<td></td>
<td></td>
<td>End of year</td>
<td></td>
</tr>
</tbody>
</table>

**Event Delivery and Participation Measures.** We collected data on the fidelity of the delivery 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.
Appendix B. Details About Data Collection and Outcome Measures

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
Appendix B. Details About Data Collection and Outcome Measures

(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 |
|---|---|---|
| Teacher survey\(^a\) | Overall | Treatment group | Control group |
| 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: \(^a\)Teacher 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.
Appendix B. Details About Data Collection and Outcome Measures

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). 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

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

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.