Providing Performance Feedback to Teachers and Principals: Implementation and Impact Findings from a Large-Scale Randomized Controlled Trial

Symposium Justification
SREE 2018 Spring Conference

Educator performance evaluation systems are a potential tool for improving student achievement by increasing the effectiveness of the educator workforce (Stecher et al., 2016; Weisburg, Daniel, Sexton, Mulhern, & Keeling, 2009). Recent research behind these performance systems encompasses the quality of the performance measures (e.g., Kane, McCaffrey, Miller, & Staiger, 2013), the implementation of such measures in performance feedback systems (e.g., Chaplin, Gill, Thompkins, & Miller, 2014), and the impact of providing performance feedback on outcomes for teachers and students (e.g., Steinberg & Sartin, 2015; Taylor & Tyler 2012).

This symposium will present findings from a large-scale randomized study that builds on the prior research base and extends our understanding of how different components of performance feedback are implemented within schools, and whether such feedback improves educator performance and student achievement.

In this study, 127 schools were randomly assigned to treatment and control conditions within eight districts. Treatment schools were provided resources and support to implement the following three performance measures in 2012-13 and 2013-14:

- a measure of teacher classroom practice with subsequent feedback sessions conducted four times per year, based on a classroom observation rubric (i.e., Classroom Assessment and Scoring System (CLASS) in four districts and Framework for Teaching (FFT) in the other four districts);

- a measure of teacher contributions to student achievement growth (i.e., value-added scores), provided to teachers and their principals once per year; and

- a measure of principal leadership based on the 360-degree Vanderbilt Assessment of Leadership in Education (VAL-ED) with subsequent feedback sessions conducted twice per year.

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. Instead, the measures were used to provide educators and their supervisors with information regarding their performance. It was hypothesized that systematic performance measurement and feedback would lead to improved teacher classroom practice, principal leadership, and ultimately, student achievement.

The four papers in this symposium test this hypothesis by addressing how different aspects of performance measurement were implemented and whether systematic feedback improved educator and student outcomes. The first paper describes the properties of the teacher classroom practice measure and how feedback based on that measure was implemented. The second paper
describes the properties of the teacher value-added measure and how feedback based on that measure was implemented. The third paper describes the properties of the principal leadership measure and how feedback based on that measure was implemented. The fourth paper addresses the question of whether feedback based on the measures affected educator performance and student achievement. Together, the four papers provide a detailed picture of what policymakers and practitioners might expect if adopting key components of an educator performance feedback system.

Dr. Andy Sokatch (Bill & Melinda Gates Foundation) has agreed to serve as the discussant.
PAPER 1

Title
Performance Measurement and Feedback: Implementation Findings for Teacher Classroom Practice

Authors
Jordan Rickles, Andrew J. Wayne, Michael S. Garet, Seth Brown, Mengli Song, and David Manzeske (American Institutes for Research)

Background
Frequent and systematic performance measurement and feedback may generate information that distinguishes between lower- and higher-performing teachers and between different dimensions of a teacher’s instructional practice, which could help identify teachers in need of support and dimensions on which a teacher should improve (see e.g., Donaldson & Papay, 2014; Papay, 2012). If the feedback is frequent and perceived as clear, fair, and useful, it may have an impact on teachers’ interest in improving along the dimensions on which they received feedback. This may lead teachers to get support for improvement, through, for example, participating in professional development activities or consulting colleagues. It may also lead teachers to independently identify and try out new classroom practices.

Purpose & Research Questions
This paper examines whether the performance measure and feedback component for teacher classroom practice, as implemented, exhibited the qualities intended for systematic and useful performance feedback. In particular, we address the following research questions:

1. To what extent was the performance measure and feedback implemented as planned?
2. To what extent did the performance measure distinguish educator performance?
3. To what extent did educators’ experiences with performance feedback differ for treatment and control schools?

Setting
To answer the study’s research questions, we recruited a sample of eight districts (spanning multiple geographic regions in the U.S.) and conducted the study in a selected group of schools in each district. The participating schools were assigned by lottery to implement the study’s intervention (the treatment group, 63 schools) or not (the control group, 64 schools). Characteristics of the participating schools are presented in Table A.1.

Participants
In the participating schools, the study focused on the teachers of mathematics and reading/English language arts (ELA) in grades 4-8. Characteristics of the participating teachers are presented in Table A.2.

Intervention
Performance feedback was based on classroom observations and was designed to provide
information on multiple dimensions of a teacher’s classroom practice repeatedly throughout the year. Specifically, it was designed to have the following features:

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

Districts were given the opportunity to choose between one of two distinct rating systems for measuring classroom practice. Four districts chose CLASS, and four districts chose FFT. The two systems capture similar dimensions of classroom practice (see Table A.3) but differ in how the observations and feedback sessions are conducted and in the amount and kind of information on teacher performance the systems’ reports provide.

**Research Design and Analysis**

The study used a multisite cluster randomized design. To address Research Questions 1 and 2, we conducted descriptive analyses of the implementation and observation ratings data collected in the treatment schools only. We used a generalizability theory framework (Shavelson and Webb 1991) to estimate the reliability of the classroom practice scores treatment teachers received. To address Research Question 3, we compared survey responses from teachers in the treatment and control groups, taking into account the clustered data structure with multilevel modeling where appropriate.

**Data Collection**

The following data were used for the analyses presented in this paper:

- **Implementation data.** 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 CLASS and FFT vendors on the frequency of classroom observations and feedback sessions.

- **Performance data.** We collected the ratings generated by the teacher classroom practice performance measures through the vendors’ online systems.

- **Experiences with performance feedback.** In the spring of each study year, we surveyed the teachers in treatment and control schools to collect information on the performance information they received.

**Findings**

The findings presented in this abstract are based on the first year of implementation. Findings from the second year of implementation are currently under review conducted by IES’s Standards and Review Office, and will be incorporated into the presentation paper once the review is complete (in fall 2017).

Key findings for Research Question 1 include the following:
• The majority of teachers were observed the intended four times and received feedback (see Figure A.1).

• The majority of the observation reports (76 percent of CLASS reports and 71 percent of FFT reports) identified at least one dimension of practice as a strength and one dimension for improvement. In addition, three quarters of the CLASS reports supported the identified dimension(s) for improvement with at least one example from the observation, but less than a quarter of the FFT reports did so.

Key findings for Research Question 2 include the following:

• For both CLASS and FFT, observation scores were concentrated at the upper end of the scale, limiting the degree of differentiation between lower- and higher-performing teachers (see Figures A.2 to A.5).

• Teachers’ overall classroom observation scores, averaged across all four windows, provided some reliable information to distinguish between lower- and higher-performing teachers, but differences in a teacher’s ratings across observations limited how much one could learn about persistent performance from a single observation (see Table A.4).

• Teachers’ overall classroom observation scores were positively, although weakly, correlated with teacher value-added scores (see Table A.5).

Key findings for Research Question 3 include the following:

• More treatment than control teachers reported having discussions about CLASS/FFT-related areas of practice with someone who provided them with performance feedback (see Figure A.6).

Conclusions

Study districts were largely successful in implementing the teacher classroom practice performance measure. However, the performance information did not fully distinguish teacher performance. In addition, for both CLASS and FFT, the scores were concentrated at the upper end of the scale and virtually all teachers had scores associated with positive performance levels.
PAPER 2

Title
Performance Measurement and Feedback: Implementation Findings for Teacher Value Added

Authors
Seth Brown, Andrew J. Wayne, Michael S. Garet, Mengli Song, Jordan Rickles, and David Manzeske (American Institutes for Research)

Background
Frequent and systematic performance measurement and feedback may generate information that distinguishes between lower- and higher-performing teachers and between different dimensions of a teacher’s instructional practice, which could help identify teachers in need of support and dimensions on which a teacher should improve (see e.g., Donaldson & Papay, 2014; Papay, 2012). Teacher value-added scores in particular are intended to differentiate teacher performance to identify lower- and higher-performing teachers, and therefore identify teachers in need of improvement. Additionally, value-added scores are intended to provide information about a teacher’s relative performance in reading/ELA versus mathematics, for those who teach both subjects. If the measures are perceived as clear, fair, and useful, it may have an impact on teachers’ perceptions of their own performance, or principals’ perceptions of their teachers’ performance. The information may lead teachers to seek out support through, for example, participating in professional development activities or consulting colleagues, or lead principals to provide the teacher with such opportunities. It may also influence teachers’ perceptions of whether they should remain in the profession, or principals’ staffing decisions.

Purpose & Research Questions
This paper examines whether the teacher value-added measure, as implemented, exhibited the qualities intended for systematic and useful performance feedback. In particular, we address the following research questions:

1. To what extent was the development and delivery of the value-added measure implemented as planned?
2. To what extent did the value-added scores distinguish educator performance?
3. To what extent did educators’ experiences with information on student achievement differ for treatment and control schools?

Setting
This paper uses the same sample of eight districts and 127 schools as the first paper in the symposium (see Table A.1).

Participants
In the participating schools, the study focused on the teachers of mathematics and reading/English language arts (ELA) in grades 4-8. Characteristics of the participating teachers are presented in Table A.2.
**Intervention**
The measure of teacher value-added was designed to provide teachers with information about their contribution to their students’ achievement growth relative to other teachers in their districts. AIR estimated individual teachers’ value-added scores with a statistical method for analyzing multiple years of students’ test score data, and compiled the results in a report provided to the teachers and teachers’ principals. Reports were designed to provide information about a teacher’s contribution to student achievement overall, and in particular grades and subjects. Each report presented a teacher’s overall value-added score, the score for each subject the teacher taught, and the score for each subject-grade combination.

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

**Research Design and Analysis**
The study used a multisite cluster randomized design. To address Research Questions 1 and 2, we conducted descriptive analyses of the implementation and value-added data collected in the treatment schools only. To address Research Question 3, we compared survey responses from teachers in the treatment and control groups, taking into account the clustered data structure with multilevel modeling where appropriate.

**Data Collection**
The following data were used for the analyses presented in this paper:

- **Implementation data.** 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 AIR’s assessment team on the teacher and principal viewing of the value-added data.

- **Performance data.** We collected value-added score information from AIR’s assessment team.

- **Experiences with performance feedback.** In the spring of each study year, we surveyed the teachers in treatment and control schools to collect information on the performance information they received.

**Findings**
The findings presented in this abstract are based on the first year of implementation. Findings from the second year of implementation are currently under review conducted by IES’s Standards and Review Office, and will be incorporated into the presentation paper once the review is complete (in fall 2017).

Key findings for Research Question 1 include the following:

- A large majority of teachers (80%) had a sufficient number of students with the achievement data required to estimate value-added scores.
• Most teachers (85%) and principals (81%) participated in the value-added report training.

• Less than half of teachers and principals accessed the reports, with access rates varying substantially across schools (see Figure A.7).

Key findings for Research Question 2 include the following:

• Treatment teachers’ value-added scores were distributed as expected (see Figure A.8).

• Many teachers with a student growth report had a value-added score that measurably differed from the district average, particularly in mathematics (see Figure A.9).

• The value-added scores provided some reliable information to distinguish between lower- and higher-performing teachers.

• Among teachers with value-added scores in both reading/English language arts (ELA) and mathematics, about half had value-added reports that suggested the teacher performed better in one subject area than the other (see Figure A.10).

Key findings for Research Question 3 include the following:

• Relative to control teachers, treatment teachers were more likely to report receiving value-added scores and less likely to report receiving test scores for individual students or classroom average scores (see Figure A.11).

• Treatment teachers were more likely than control teachers to perceive the student achievement information they received as difficult to understand, but they were more likely to perceive it as fair (see Figure A.12).

Conclusions
Study districts were successful at implementing the study’s teacher value-added measure in some respects. On the one hand, they successfully supplied the data needed, and value-added scores were computed for a large majority of teachers. On the other hand, although most teachers and principals attended training webinars prior to the release of the reports, fewer than half of teachers and principals accessed their reports. Meanwhile, the value-added measure distinguished teacher performance but only to a degree. Treatment teachers received more feedback on student achievement. However, among treatment and control teachers who received student achievement data, treatment teachers reported more negative perceptions about the information they received on their students’ achievement.
PAPER 3

Title
Performance Measurement and Feedback: Implementation Findings for Principal Leadership

Authors
Andrew J. Wayne, Michael S. Garet, Seth Brown, Mengli Song, Jordan Rickles, and David Manzeske (American Institutes for Research)

Background
Frequent and systematic performance measurement and feedback may generate information that distinguishes between lower- and higher-performing principals and between different dimensions of a principal’s leadership practices, which could help identify principals in need of support and dimensions on which a principal should improve. If the feedback is frequent and perceived as clear, fair, and useful, it may have an impact on principals’ interest in improving along the dimensions on which they received feedback. This may lead principals to get support for improvement, through, for example, participating in professional development activities or consulting colleagues. It may also lead principals to independently identify and try out new leadership practices.

Purpose & Research Questions
This paper examines whether the performance measure and feedback component for principal leadership, as implemented, exhibited the qualities intended for systematic and useful performance feedback. In particular, we address the following research questions:

1. To what extent was the performance measure and feedback implemented as planned?
2. To what extent did the performance measure distinguish principal performance?
3. To what extent did principals’ experiences with performance feedback differ for treatment and control schools?

Setting
This paper uses the same sample of eight districts and 127 schools as the first paper in the symposium (see Table A.1).

Participants
The study focused on the principals of the participating schools, which included elementary schools and middle schools. Characteristics of the participating principals are presented in Table A.6.

Intervention
Feedback on principal leadership was based on the Vanderbilt Assessment of Leadership in Education (VAL-ED), a 360-degree survey assessment administered twice a year to principals, principal supervisors, and teachers. The VAL-ED includes six “core components” of principal performance and six “key processes” (see Table A.7). A report for each principal was generated
after each administration of the VAL-ED (fall and spring), and the principal’s supervisor was expected to discuss the report with the principal in a feedback session.

The VAL-ED reports present scores and performance levels, as well as percentile ranks, for each dimension of leadership. VAL-ED reports present an overall score, a score for each core component, and a score for each key process. For each of these 13 scores, the report additionally presents a performance label and a percentile rank, relative to the principals included in a national VAL-ED field test. Each score is an average across the three respondent groups (i.e., principal, supervisor, and teachers), with each group weighted equally. The report additionally shows the scores received from each respondent group separately.

**Research Design and Analysis**

The study used a multisite cluster randomized design. To address Research Questions 1 and 2, we conducted descriptive analyses of the implementation and leadership ratings data collected in the treatment schools only. We used a generalizability theory framework (Shavelson and Webb 1991) to estimate the reliability of the leadership scores treatment principals received. To address Research Question 3, we compared survey responses from principals in the treatment and control groups, taking into account the clustered data structure with multilevel modeling where appropriate.

**Data Collection**

The following data were used for the analyses presented in this paper:

- **Implementation data.** We documented attendance at orientation and training events related to the study’s performance measures. We also gathered data from the online system maintained by the VAL-ED vendor on the response rates of teachers, principals, and principal supervisors to the VAL-ED surveys. Finally, we collected logs that principal supervisors completed about each feedback session they conducted.

- **Performance data.** We collected the ratings generated by the VAL-ED through the vendors’ online systems, including ratings by teachers, principals, and principal supervisors.

- **Experiences with performance feedback.** In the spring of each study year, we surveyed the principals in treatment and control schools to collect information on the performance information they received.

**Findings**

The findings presented in this abstract are based on the first year of implementation. Findings from the second year of implementation are currently under review conducted by IES’s Standards and Review Office, and will be incorporated into the presentation paper once the review is complete (in fall 2017).

Key findings for Research Question 1 include the following:
• All principals and their supervisors completed the VAL-ED rating form, and a high percentage of teachers in each treatment school (80 percent in fall and 90 percent in spring on average) also completed the form.

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

Key findings for Research Question 2 include the following:

• In the fall and spring, principals’ overall scores were distributed across the four performance levels (see Figure A.13).

• VAL-ED ratings provided by principals, supervisors, and teachers in the fall were often too different to form a reliable measure, but the spring ratings were consistent enough to distinguish between some lower- and higher-performing principals. In the fall, agreement among the three respondent groups' overall scores was low, with correlations ranging from .06 to .27. In the spring, correlations were higher (between .26 and .38), and thus the reports provided a more consistent message about a principal’s effectiveness (see Table A.8).

Key findings for Research Question 3 include the following:

• Treatment principals reported receiving more feedback than control principals (see Figure A.14).

• Among those who reported receiving feedback, most principals in both treatment and control schools had positive perceptions about the feedback they received (see Figure A.15).

Conclusions
Study districts were successful in implementing the principal leadership performance measure. Principals and their supervisors were trained for their roles, and all of the planned feedback sessions occurred. The VAL-ED provided performance information that categorized principals as lower- or higher-performing. However, differences in ratings across survey respondent groups limited the consistency of the performance information provided.
Title
Providing Performance Feedback to Teachers and Principals: Impact on Educator Practice and Student Achievement

Authors
Mengli Song, Michael S. Garet, Andrew J. Wayne, Seth Brown, Jordan Rickles, and David Manzeske (American Institutes for Research)

Background
Educator performance evaluation systems are a potential tool for improving student achievement by increasing the effectiveness of the educator workforce (Stecher et al., 2016; Weisburg, Daniel, Sexton, Mulhern, & Keeling, 2009). 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, which may in turn lead to improved student achievement.

Purpose & Research Questions
This paper presents findings about the impact of an intervention designed to provide performance feedback to teachers and principals based on a large-scale randomized controlled trial (RCT). Specifically, it addresses the following research questions:

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

Intervention
The intervention consisted of the following three types of performance measures that were implemented during the 2012–13 (i.e., Year 1) and 2013–14 (i.e., Year 2) school years:

- **Classroom practice measure**: a measure of teacher classroom practice with subsequent feedback sessions conducted four times per year, based on a classroom observation rubric (i.e., CLASS in four districts and FFT in the other four districts);

- **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; and

- **Principal leadership measure**: a measure of principal leadership based on a 360-degree survey (VAL-ED) with subsequent feedback sessions conducted twice per year.

Setting
The RCT took place in 127 regular elementary and middle schools in eight purposively selected...
districts where existing policies for the evaluation of teachers and principals differed substantially from the study’s intervention. The eight districts were located in five states and spanned all geographic regions except the Northeast, with two or three districts in each region. The sample was mostly urban, with only one suburban and one rural district.

Participants
Participants of the study included the principals and teachers of mathematics and reading/English language arts (ELA) in grades 4–8 from the study schools. The study sample also included students in grades 4–8 who were present in the study schools near the end of each of the two intervention years.

Research Design
The study used a multisite cluster randomized design. To maximize the precision of impact estimates, random assignment of schools was conducted separately within each of the 37 blocks in the eight study districts. The blocks were defined by district and school level (elementary schools or middle schools), and 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. Both groups continued to implement their district’s existing educator evaluation systems, but the treatment schools also implemented the intervention.

Data Collection
The following types of data on teacher, principal, and student outcomes were collected and used for the analyses presented in this paper:

- **Teacher classroom practice.** 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.\(^1\) We coded each of the videos using both the CLASS and the FFT, which allowed us to examine the intervention’s impact (1) on a measure of practice aligned with the measure selected for feedback and (2) on a measure that was similar but not completely aligned.

- **Principal leadership.** 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).

- **Student achievement.** To measure student achievement, we collected students’ scores on state standardized tests in reading/ELA and mathematics. To form common metrics of student achievement 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.

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

\(^1\) We video recorded two lessons for some teachers and one for others to achieve the desired precision while minimizing cost and burden.
Analytic Methods
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”). Based on the impact samples, we assessed the impacts of the study’s intervention on different types of outcomes using different analytic models, as summarized below.

- To assess the impact on teacher classroom practice, we used observation data to estimate a three-level model (with lessons nested within teachers and teachers nested within schools).

- 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 estimated a three-level model (where students are nested within teachers and teachers nested within schools) with data pooled across grades 4–8.

All impact models incorporated fixed effects for random assignment blocks as well as 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.

Findings
The study’s final report that includes detailed findings about the impact of the study’s intervention on educator practice and student achievement is currently in the final stage of review by IES’s Standards and Review Office. We will be able to share the findings once the report is released, which is scheduled for fall 2017, well before the spring 2018 SREE conference.
### Table A.1. School background characteristics, by study group

<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 and 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><strong>Number of schools</strong></td>
<td><strong>63</strong></td>
<td><strong>64</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**NOTE:** The analyses are based on an OLS regression model controlling for random assignment blocks. The treatment group means are unadjusted means; the control group means were computed by subtracting the estimated group differences from the unadjusted treatment group means. The p values are based on t tests. Two-tailed statistical significance at the p < .05 level is indicated by an asterisk (*).  
**SOURCE:** 2011–12 Common Core of Data.

### Table A.2. Teacher background characteristics, fall 2012, by study group (grades 4–8)

<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>
<tr>
<td><strong>Number of teachers</strong></td>
<td><strong>575</strong></td>
<td><strong>594</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**NOTE:** The analyses are based on a two-level linear regression model controlling for random assignment blocks. The treatment group means are unadjusted means, and the control group means were computed by subtracting the estimated group differences from the unadjusted treatment group means. The p values are based on t tests. Two-tailed statistical significance at the p < .05 level is indicated by an asterisk (*).  
**SOURCE:** Fall 2012 District Archival Records.
Table A.3. Domains and dimensions of classroom practice for CLASS and FFT

<table>
<thead>
<tr>
<th>Classroom Assessment 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>

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.
Table A.4. Summary of reliability estimates for classroom observation overall scores

<table>
<thead>
<tr>
<th>Measure</th>
<th>Reliability Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classroom observation overall scores</td>
<td></td>
</tr>
<tr>
<td>CLASS single-window score</td>
<td>.24</td>
</tr>
<tr>
<td>FFT single-window score</td>
<td>.49</td>
</tr>
<tr>
<td>CLASS four-window average score</td>
<td>.42 to .50</td>
</tr>
<tr>
<td>FFT four-window average score</td>
<td>.69 to .75</td>
</tr>
</tbody>
</table>

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

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

Table A.5. Pairwise correlations between classroom observation overall scores and prior-year value-added scores

<table>
<thead>
<tr>
<th>Measure</th>
<th>Overall Correlation Coefficient</th>
<th>Mathematics Correlation Coefficient</th>
<th>ELA/Reading Correlation Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>CLASS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Four-window average</td>
<td>253</td>
<td>.09</td>
<td>198</td>
</tr>
<tr>
<td>Window 1</td>
<td>217</td>
<td>.07</td>
<td>170</td>
</tr>
<tr>
<td>Window 2</td>
<td>251</td>
<td>.11</td>
<td>196</td>
</tr>
<tr>
<td>Window 3</td>
<td>252</td>
<td>.11</td>
<td>197</td>
</tr>
<tr>
<td>Window 4</td>
<td>226</td>
<td>.00</td>
<td>186</td>
</tr>
<tr>
<td>FFT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Four-window average</td>
<td>173</td>
<td>.17*</td>
<td>142</td>
</tr>
<tr>
<td>Window 1</td>
<td>169</td>
<td>.15</td>
<td>138</td>
</tr>
<tr>
<td>Window 2</td>
<td>171</td>
<td>.09</td>
<td>140</td>
</tr>
<tr>
<td>Window 3</td>
<td>173</td>
<td>.17*</td>
<td>142</td>
</tr>
<tr>
<td>Window 4</td>
<td>171</td>
<td>.10</td>
<td>141</td>
</tr>
</tbody>
</table>

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

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

**SOURCE:** Teachstone Online System (CLASS), Teachscape Online System (FFT), and Student Growth Reporting System.
### Table A.6. Principal background characteristics, fall 2012, by study group

<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

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

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

**SOURCE**: Fall 2012 District Archival Records.

### Table A.7. VAL-ED core components and key processes

<table>
<thead>
<tr>
<th>Core components</th>
<th>Key processes</th>
</tr>
</thead>
<tbody>
<tr>
<td>High standards for student learning</td>
<td>Planning</td>
</tr>
<tr>
<td>Rigorous curriculum</td>
<td>Implementing</td>
</tr>
<tr>
<td>Quality instruction</td>
<td>Supporting</td>
</tr>
<tr>
<td>Culture of learning and professional behavior</td>
<td>Advocating</td>
</tr>
<tr>
<td>Connections to external communities</td>
<td>Communicating</td>
</tr>
<tr>
<td>Systemic performance accountability</td>
<td>Monitoring</td>
</tr>
</tbody>
</table>

### Table A.8. Correlations between VAL-ED respondent group overall scores from different respondent groups in fall and spring

<table>
<thead>
<tr>
<th>Correlation</th>
<th>Fall 2012</th>
<th>Spring 2013</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>

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

* Significantly different from zero with p < .05.

**SOURCE**: Fall 2012 and Spring 2013 VAL-ED Surveys.
Figure A.1. Percentage of teachers who received one, two, three, or four study observations and feedback sessions in CLASS and FFT districts

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

*SOURCE:* Teachstone Online System and Teachscape Online System.
Figure A.2. Distribution of teachers across performance levels based on CLASS overall scores, by observation window and the four-window average

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

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

SOURCE: Teachstone Online System.
Figure A.3. Distribution of teachers based on their CLASS overall scores in each observation window and the four-window average

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

SOURCE: Teachstone Online System.
Figure A.4. Distribution of teachers across study-defined performance levels based on FFT overall scores, by observation window and the four-window average

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

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

SOURCE: Teachscape Online System.
Figure A.5. Distribution of teachers based on their FFT overall scores in each observation window and the four-window average

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

SOURCE: Teachscape Online System.
Figure A.6. Percentage of teachers who reported discussing areas of classroom practice related to CLASS/FFT and areas not related, with someone who provided them with feedback during the school year, by treatment status.

Areas of classroom practice related to CLASS/FFT

- Behavior management: 56% (Treatment), 51% (Control)
- Classroom organization: 52%* (Treatment), 40% (Control)
- Emotional support for students: 50%* (Treatment), 39% (Control)
- Instructional dialogue: 72%* (Treatment), 72% (Control)
- Student engagement: 54% (Treatment), 74%* (Control)

Areas of classroom practice not related to CLASS/FFT

- Lesson planning: 47% (Treatment), 49% (Control)
- Data use: 58% (Treatment), 62% (Control)
- Content-specific teaching techniques: 51% (Treatment), 53% (Control)
- Content knowledge: 48% (Treatment), 51% (Control)

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

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

Figure A.7. Value-added report access rates for teachers and principals, by district

NOTE: Sample size = 433 teachers and 62 schools.
SOURCE: AIR value-added system.
Figure A.8. Distribution of treatment teachers based on their value-added scores

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

SOURCE: AIR value-added system.
Figure A.9. Distribution of treatment teachers based on whether their value-added score was considered measurably above or below the district average, overall and by subject

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

**SOURCE:** AIR value-added system.
Figure A.10. Distribution of treatment teachers based on their subject area value-added scores being considered measurably above or below the district average

<table>
<thead>
<tr>
<th>Reading/ELA score</th>
<th>Mathematics score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Measurably below average</td>
</tr>
<tr>
<td>Measurably below average</td>
<td>7.1%</td>
</tr>
<tr>
<td>Not measurably different from average</td>
<td>17.2%</td>
</tr>
<tr>
<td>Measurably above average</td>
<td>0.8%</td>
</tr>
</tbody>
</table>

NOTE: The distribution of teachers is based on whether the 80 percent confidence interval for a teacher’s value-added score in reading/ELA and mathematics was above or below the district average. Sample size = 239 teachers.

SOURCE: AIR value-added system
Figure A.11. Percentage of teachers who reported receiving specific types of student achievement information, by treatment status

<table>
<thead>
<tr>
<th>Type of Information</th>
<th>Treatment</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value-added scores based on students that I taught</td>
<td>45%</td>
<td>24%</td>
</tr>
<tr>
<td>Data on individual students that I taught</td>
<td>64%*</td>
<td>84%</td>
</tr>
<tr>
<td>Average data for classes of students that I taught</td>
<td>51%*</td>
<td>62%</td>
</tr>
</tbody>
</table>

% of teachers who reported receiving student achievement information

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

Figure A.12. Percentage of teachers receiving student achievement information who agreed or strongly agreed with statements about that information, by treatment status.

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

Figure A.13. Distribution of treatment principals across performance levels based on VAL-ED overall scores in fall and spring

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

SOURCE: Fall 2012 and Spring 2013 VAL-ED Surveys.
Figure A.14. Number of feedback instances and duration of oral feedback that principals reported receiving, by treatment status

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

SOURCE: Spring 2013 Principal Survey.
Figure A.15. Percentage of principals receiving performance feedback who agreed or strongly agreed with statements about that feedback, by treatment status

<table>
<thead>
<tr>
<th>Statement</th>
<th>Treatment</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feedback was a fair assessment of my performance</td>
<td>86%</td>
<td>71%</td>
</tr>
<tr>
<td>Feedback included specific ideas about how I could improve my performance</td>
<td>80%</td>
<td>66%</td>
</tr>
</tbody>
</table>

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

SOURCE: Spring 2013 Principal Survey.
SYMPOSIUM REFERENCES


