December 2014

Making Connections

Measuring school leaders' effectiveness: An interim report from a multiyear pilot of Pennsylvania's Framework for Leadership

Bing-ru Teh
Hanley Chiang
Stephen Lipscomb
Brian Gill
Mathematica Policy Research

Key findings

- School leaders who earned higher scores in one category of leadership practices measured by the Framework for Leadership tended to earn higher scores in the other categories.
- On each measured leadership practice most school leaders earned scores in the top two of four possible performance levels.
- School leaders with larger estimated contributions to student achievement growth did not score higher than school leaders with smaller estimated contributions.

ies NATIONAL CENTER FOR EDUCATION EVALUATION AND REGIONAL ASSISTANCE
Institute of Education Sciences
U.S. Department of Education

REL MID-ATLANTIC
Regional Educational Laboratory
At ICF International

REL 2015–058

The National Center for Education Evaluation and Regional Assistance (NCEE) conducts unbiased large-scale evaluations of education programs and practices supported by federal funds; provides research-based technical assistance to educators and policymakers; and supports the synthesis and the widespread dissemination of the results of research and evaluation throughout the United States.

December 2014

This report was prepared for the Institute of Education Sciences (IES) under Contract ED-IES-12-C-0006 by Regional Educational Laboratory Mid-Atlantic administered by ICF International. The content of the publication does not necessarily reflect the views or policies of IES or the U.S. Department of Education nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

Teh, B., Chiang, H., Lipscomb, S., & Gill, B. (2014). *Measuring school leaders' effectiveness: An interim report from a multiyear pilot of Pennsylvania's Framework for Leadership* (REL 2015–058). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Mid-Atlantic. Retrieved from http://ies.ed.gov/ncee/edlabs.

This report is available on the Regional Educational Laboratory website at http://ies.ed.gov/ncee/edlabs.

Summary

States and districts across the country are revising how they evaluate school principals. Since 2012, 42 states and the District of Columbia have received waivers from particular requirements of the No Child Left Behind Act in return for committing to several reforms, including developing new systems for evaluating principals. Unfortunately, there is scant evidence on the accuracy of current evaluation tools.

Pennsylvania is among the states that must develop a new tool for evaluating principals and assistant principals (collectively referred to as school leaders). Starting in 2014/15, half of a school leader's annual evaluation rating must be based on a supervisor's assessment of the quality of leadership practices. The remaining half must be based on measures of student achievement.

The Pennsylvania Department of Education developed an evaluation tool called the Framework for Leadership (FFL), which rates school leaders in 19 leadership practices as distinguished, proficient, needs improvement, or failing. The practices are grouped into four categories: strategic/cultural leadership, systems leadership, leadership for learning, and professional and community leadership. The evaluation tool was piloted in 2012/13 and 2013/14 on selected groups of school leaders, in preparation for introducing it statewide in 2014/15.

The Pennsylvania Department of Education, a member of the Principal Evaluation Research Alliance of the Regional Education Laboratory Mid-Atlantic, requested statistical evidence on how well FFL scores measure school leaders' effectiveness. This study uses data from the 2012/13 pilot evaluations to analyze three key FFL properties:

- Internal consistency—the degree to which different parts of the FFL come to similar conclusions about a school leader's effectiveness. This is desirable because the leadership qualities captured by different parts of the FFL are supposed to reflect an overall capability to improve student achievement through effective school leadership.
- Score variation—the degree to which scores differ across school leaders. Score variation is necessary for the FFL to differentiate between high- and low-performing school leaders, a basic goal for any evaluation tool.
- Concurrent validity—the degree to which FFL scores in a given year correlate with school leaders' contributions to student achievement growth in the same year, indicating that FFL scores reflect leadership practices that contribute to raising student achievement.

Key findings

This interim report provides findings and considerations based on the pilot evaluation data from 2012/13 for 336 principals and 69 assistant principals in Pennsylvania:

- The full FFL had good internal consistency for both principals and assistant principals. School leaders who earned higher scores in one category of leadership practices tended to earn higher scores in the other categories.
- Most school leaders received scores of proficient or distinguished for specific leadership practices. Supervisors rated the performance of both principals and assistant principals as proficient or distinguished 95 percent of the time and as needing

improvement in the remaining 5 percent. The most common rating was proficient (70 percent for principals and 79 percent for assistant principals). Supervisors rarely assigned a failing rating: only two principals received a failing rating on a component, while no assistant principal received a failing rating.

- School leaders with larger estimated contributions to student achievement growth did not, on average, receive higher FFL scores than school leaders with smaller estimated contributions to student achievement growth.

Interim conclusions and suggestions

The findings from the 2012/13 pilot reveal both strengths and weaknesses of the FFL. The good internal consistency of the full FFL suggests that it is based on a coherent definition of leadership quality. However, the concentration of FFL scores in the top two performance levels contrasts with prior research that has revealed clear differences in the contributions principals make to student achievement growth. Supervisors may thus have rated their school leaders too positively. This possibility is substantiated by the absence of a positive correlation between school leaders' FFL scores and their contributions to student achievement growth.

This lack of correlation between FFL scores and school leaders' contributions to student achievement growth does not necessarily make FFL scores a less valid measure of school leaders' effectiveness than scores from other tools. To date, there is no robust evidence that any current school leader evaluation tool is associated with school leaders' contributions to student achievement growth.

Nevertheless, these findings suggest that more evidence is needed on the validity of using FFL scores to identify effective and ineffective school leaders. The Pennsylvania Department of Education may need to consider additional measures of school leaders' performance, such as anonymous ratings by teachers, that may be less susceptible to excessive leniency. Even if the additional measures do not factor officially into evaluations, they can be compared with FFL scores as a check on whether supervisors are being too lenient in assigning ratings in the FFL.

More specific guidance on how to determine ratings would also help supervisors determine whether they are rating school leaders appropriately. Providing examples of evidence that would merit each possible score for every FFL component would give supervisors a benchmark for their evaluations of school leaders.

Contents

Box

Figures

Tables

Why this study?

States and districts across the country are revising how they evaluate school principals. Since 2012, 42 states and the District of Columbia have received waivers from particular requirements of the No Child Left Behind Act in return for committing to several reforms, including developing and implementing new systems for evaluating principals that take into account student achievement growth and the quality of principals' leadership practices.

Need for accurate evaluation tools

A key task for states and districts that are revising their systems for evaluating principals is to select or develop accurate evaluation tools. Unfortunately, there is scant evidence on the accuracy of current evaluation tools. A recent review found that 63 of 65 principal evaluation tools had no documented reliability or validity (Goldring et al., 2009). No evaluation tool has been consistently shown to indicate principals' contributions to student achievement, even though improving student outcomes is a central task of school leaders (see appendix A for a more extensive discussion of the literature on the effectiveness of school principals). There is a substantial need for more evidence on ways to accurately measure the quality of principals' leadership practices.

A key task for states and districts that are revising their systems for evaluating principals is to select or develop accurate evaluation tools. Unfortunately, there is scant evidence on the accuracy of current evaluation tools

Pennsylvania is among the states that must develop and implement a new tool for evaluating principals and assistant principals (collectively referred to as school leaders). Under 2012 legislation half of a school leader's annual evaluation rating must be based on a supervisor's assessment of the quality of leadership practices. The remaining half must be based on measures of student achievement.[1] Beginning in 2014/15, this new evaluation system will apply to all school leaders in the state.

Pennsylvania's Framework for Leadership

The Pennsylvania Department of Education developed an evaluation tool called the Framework for Leadership (FFL) to measure the quality of school leaders' practices. The FFL specifies 19 leadership practices, known as components, on which each school leader is rated by an administrator who has supervisory authority over the school leader, such as a superintendent or assistant superintendent. On each component a school leader can receive a rating of distinguished (3 points), proficient (2 points), needs improvement (1 point), or failing (0 points).

FFL components are grouped into four domains: strategic/cultural leadership, systems leadership, leadership for learning, and professional and community leadership (see appendix B for a list of components grouped by domain). For each domain a school leader's supervisor is supposed to judge the preponderance of evidence from the components in the domain to assign a summary score, known as a domain score, using the same rating scale as for the component scores (3, 2, 1, or 0 points). Supervisor ratings are based on direct observation and on evidence submitted by the school leaders.

Because there has been little research on how accurately tools such as the FFL measure school leaders' performance, the Pennsylvania Department of Education (a member of the Principal Evaluation Research Alliance of the Regional Educational Laboratory

Mid-Atlantic) requested statistical evidence on how well FFL scores measure school leaders' effectiveness. In particular, the Pennsylvania Department of Education asked for evidence on three key FFL properties:

- Internal consistency, a reliability measure capturing the degree of consistency in the same leader's scores from different parts of the FFL.
- Score variation, the degree of score differences across school leaders, which determines whether the FFL can distinguish high and low performers.
- Concurrent validity, the degree to which the FFL measures the concept it is intended to measure—school leaders' effectiveness in raising student achievement.

Examining FFL properties can help Pennsylvania stakeholders refine or modify the tool to improve its accuracy. In addition, evidence on the FFL's strengths and weaknesses can help other states and districts that are developing or refining their own tools for measuring school leaders' effectiveness.

Examining the properties of the state's leadership evaluation tool can help Pennsylvania stakeholders refine or modify the tool to improve its accuracy

Partly to collect evidence on FFL properties, the Pennsylvania Department of Education piloted the evaluation tool with selected groups of school leaders before introducing the tool statewide. The FFL pilots occurred in the 2012/13 and 2013/14 school years; the statewide rollout will begin in 2014/15 (see appendix C for a description of the participants, rating procedures, and completeness of data in the 2012/13 pilot year). The pilot evaluations were used only to provide evidence on FFL properties. This interim report provides findings based on data from the 2012/13 pilot year.

What the study examined

Using data from the 2012/13 pilot year, this study sought to characterize the FFL's internal consistency, its score variation, and the relationship of its scores with school leaders' contributions to student achievement growth. The first two properties were examined using descriptive analyses, and the third using correlational analyses (see box 1 for an overview of the study's data and methods and appendixes C–G for more detail).

Descriptive research questions

What is the internal consistency of the full FFL and its domains?

Internal consistency—the degree to which different parts of the FFL come to similar conclusions about a school leader's effectiveness—is desirable because the leadership qualities captured by different parts of the FFL are supposed to reflect an overall capability to improve student achievement through effective school leadership. The evaluation tool is based on a common conception of effective school leadership, so the same leader's scores on different parts of the FFL should be consistent.

Internal consistency is the only type of reliability the study can examine with the pilot evaluation data. Because each school leader is rated by only one supervisor and only once in each pilot year, the study cannot examine the degree of consistency in a leader's FFL scores from different supervisors (inter-rater reliability) or across different but close points in time (test-retest reliability).

Box 1. Data and methods

Data

The data for the study consisted of school leaders' scores on the Framework for Leadership (FFL), school leaders' job assignments and background characteristics, and student achievement scores and background characteristics (see appendix C for a detailed description of each data source).

The study used FFL scores from the end of the 2012/13 pilot year for 336 principals and 69 assistant principals. Participating school leaders work primarily in districts receiving U.S. Department of Education Race to the Top funds and so do not necessarily represent Pennsylvania's population of school leaders. School leaders decided jointly with their supervisors which FFL components to use in the pilot evaluations, but all school leaders included in the analyses were rated on at least one component from every domain and on an average of 16 of 19 components. Although actual FFL evaluations starting in 2014/15 will require supervisors to assign a domain score based on the preponderance of evidence within a domain, supervisors in the 2012/13 pilot evaluations assigned only component scores. For the analysis, the study team computed a school leader's domain score as the equal-weighted average of scores from the components on which the leader was evaluated in that domain. The Pennsylvania Department of Education regards the four domains as equally weighted elements of a school leader's annual evaluation rating, so the study defined a school leader's full FFL score as the equal-weighted average of the four domain scores.

Data on school leaders' job assignments and background characteristics linked principals and assistant principals to the schools they led, enabling the study to attribute achievement growth at those schools to the school leaders. The data covered all Pennsylvania principals and assistant principals from 2007/08 to 2012/13.

Data on student achievement scores and background characteristics enabled the study to estimate school leaders' contributions to achievement growth that controlled for students' prior achievement and backgrounds. The data covered all Pennsylvania students in grades 3–12, with achievement data available from 2006/07 to 2012/13 and other background data available from 2007/08 to 2012/13. The student achievement growth data included scores from end-of-grade assessments (the Pennsylvania System of School Assessment), administered in grades 3–8 and 11, and end-of-course assessments (the Keystone Exams), administered primarily in high school.

Methods

Analyses to address the research question on internal consistency used data on FFL scores to calculate Cronbach's α, a measure of internal consistency that ranges from 0 to 1 (Cronbach 1951; see appendix D for a detailed discussion). The study calculated Cronbach's α for the full FFL and for each of the four domains.

Analyses to address the research question on score variation described the distributions of FFL scores on each component, each domain, and the full FFL. The distribution of scores on a component was characterized by the percentages of school leaders who earned each of the four possible scores (distinguished, proficient, needs improvement, and failing) on the component. Differences in average scores across components reflected differences in the difficulty of scoring well on those components (see appendix E for technical details). The distributions of scores on each domain and the full FFL were characterized by the percentages of school leaders in different intervals of the 0–3 point scale.

(continued)

Box 1. Data and methods *(continued)*

Analyses to address the research question on concurrent validity used student achievement and background data to estimate school leaders' contributions to student achievement growth in 2012/13, referred to as the leaders' value-added. The study estimated leaders' value-added in one of two ways. For recently hired school leaders, defined as leaders who began their current leadership roles in 2008/09 or later, value-added was estimated as the school's contribution to student achievement growth in 2012/13, adjusted for the same school's contribution under the current leader's predecessor. For longer-serving school leaders, defined as leaders who began their current leadership roles before 2008/09, value-added was estimated as the school's contribution to student achievement growth in 2012/13 without further adjustment because achievement growth data for their predecessors were not available (see appendix F for technical details on value-added estimation). The final step was to estimate a regression model for the relationship between school leaders' FFL scores from the end of the 2012/13 school year and their estimated value-added in the same year (see appendix G for technical details on this model).

To what extent do scores on the full FFL, its domains, and its components vary across school leaders?

The degree of variation in FFL scores is one indication of how well the evaluation tool differentiates between high- and low-performing school leaders. Prior research has revealed clear differences in principals' effectiveness in raising student achievement (Branch, Hanushek, & Rivkin, 2012; Chiang, Lipscomb, & Gill, 2012; Coelli & Green, 2012; Dhuey & Smith, 2012a, 2012b). To distinguish between high and low performers, FFL scores should thus also differ meaningfully.

Correlational research question

To what extent do school leaders' FFL scores correlate with their contributions to student achievement growth?

This third property indicates the validity of using FFL scores to distinguish effective and ineffective school leaders. The FFL aims to measure the leadership qualities needed to improve student achievement. Therefore, school leaders with larger contributions to achievement should receive higher FFL scores. The study assessed the FFL's concurrent validity by comparing school leaders' FFL scores with a measure of their contributions to student achievement growth on statewide assessments in the same year.[2]

What the study found

This section describes the findings on the three key properties of the FFL: its internal consistency, its score variation, and the relationship of its scores with school leaders' contributions to student achievement growth.

The full Framework for Leadership had good internal consistency for both principals and assistant principals

Internal consistency provides some assurance that an evaluation tool measures a coherent conception of performance. School leaders who score well on a particular FFL component should score well on other components in the same domain because they all describe the same dimension of leader effectiveness. If that is not the case, either the components are not grouped appropriately or the domain-level concept they are trying to describe needs refinement. Similarly, school leaders who score well in one FFL domain should score well in other domains because all the domains describe the underlying capability of a leader to raise student achievement through effective school leadership.

The standard measure of internal consistency is Cronbach's alpha (α), a statistic that ranges from 0 to 1, where larger values are associated with higher internal consistency. (The formula for Cronbach's α is provided in appendix D.) The following critical α values are used in this study:

- 0.8 or higher is considered good.
- 0.7 or higher but less than 0.8 is considered acceptable.
- 0.6 or higher but less than 0.7 is considered marginally acceptable.
- Below 0.6 is considered not acceptable.

Different domains of the Framework for Leadership yielded similar assessments of a school leader's effectiveness

Prior analyses of the Framework for Teaching in Pennsylvania have used the same critical values for good and acceptable internal consistency (Walsh & Lipscomb, 2013), which come from a textbook on surveys in social research (de Vaus 2002). Specific guidelines pertaining to evaluation tools for teachers and school leaders are not available. This study adopts an additional critical value to indicate marginally acceptable internal consistency because 0.7 is not a strict threshold for whether an evaluation tool should be implemented.

The values of Cronbach's α for principals and assistant principals indicate that the full FFL had good internal consistency for both types of school leaders (table 1), implying that the FFL's different domains yielded similar assessments of a school leader's effectiveness. The level of internal consistency corresponds closely with that of the full Framework for Teaching, which Pennsylvania is using for teacher evaluations (Walsh and Lipscomb, 2013).

The internal consistency of Framework for Leadership domains was higher for principals than for assistant principals. The internal consistency of FFL domains, which captures the similarity of a school leader's scores on components in the same domain, was uniformly higher for principals than for assistant principals (table 2). For principals, internal

Table 1. The full Framework for Leadership had good internal consistency in the 2012/13 pilot year

School leader type	Internal consistency (Cronbach's α)	Sample size
Principals	0.88	336
Assistant principals	0.85	69

Note: 0.8 or higher is good; 0.7 or higher but less than 0.8 is acceptable; 0.6 or higher but less than 0.7 is marginally acceptable; below 0.6 is not acceptable.

Source: Authors' calculations based on Framework for Leadership 2012/13 pilot evaluation scores provided by the Pennsylvania Department of Education.

consistency was acceptable for domains 1 (strategic/cultural leadership) and 2 (systems leadership), good for domain 3 (leadership for learning), and marginally acceptable for domain 4 (professional and community leadership). For assistant principals, internal consistency was marginally acceptable for domains 1–3 and not acceptable for domain 4.[3]

The findings on the internal consistency of FFL domains provide some assurance against the concern that allowing supervisors and school leaders to choose which components to use in evaluations—as they did in 2012/13—will distort FFL scores. A benefit of an internally consistent measure is that conclusions are less sensitive to which parts of the measure are used or excluded (provided that it is not substantially more difficult to be rated well on some components than others). However, to incorporate the greatest amount of information into ratings, supervisors should use as many components as they can.

The current set of components in the professional and community leadership domain exhibited the weakest relationship to each other for both types of school leaders

This study cannot determine why internal consistency of the domains was higher for principals than for assistant principals. One possible explanation is that superintendents and assistant superintendents, who supplied most of the ratings for both principals and assistant principals (figure C1 in appendix C), had less direct knowledge about assistant principals' performance. If so, component scores for assistant principals would be subject to more error and consequently would be less consistent. Another possible explanation is that supervisors may have rated assistant principals on some components that were not part of the assistant principals' responsibilities, so scores on those components would not be closely related to scores on components pertaining to the assistant principals' responsibilities.

Internal consistency was lowest in domain 4 (professional and community leadership) for both types of school leaders, especially for assistant principals. The internal consistency findings for domain 4, which measures professional and community leadership, suggest that the domain may need further development. The current set of components in the domain exhibited the weakest relationship to each other for both types of school leaders (see table 2). Domain 4 was the only one where Cronbach's α fell into the marginally acceptable range for principals and far below marginally acceptable for assistant principals (0.20).

Table 2. The internal consistency of Framework for Leadership domains was higher for principals than for assistant principals in the 2012/13 pilot year

School leader type and Framework for Leadership domain	Internal consistency (Cronbach's α)	Sample size
Principals		
Domain 1: Strategic/cultural leadership	0.79	252
Domain 2: Systems leadership	0.78	248
Domain 3: Leadership for learning	0.82	254
Domain 4: Professional and community leadership	0.68	259
Assistant principals		
Domain 1: Strategic/cultural leadership	0.62	54
Domain 2: Systems leadership	0.67	51
Domain 3: Leadership for learning	0.65	53
Domain 4: Professional and community leadership	0.20	56

Note: 0.8 or higher is good; 0.7 or higher but less than 0.8 is acceptable; 0.6 or higher but less than 0.7 is marginally acceptable; below 0.6 is not acceptable.

Source: Authors' calculations based on Framework for Leadership 2012/13 pilot evaluation scores provided by the Pennsylvania Department of Education.

The study could not determine why internal consistency was lowest in domain 4. One possibility is that the domain might not have enough components. Currently, it includes three components, while the other domains have five or six. Adding components to a scale measure typically increases the scale's internal consistency by incorporating more information on the underlying concept of interest. Another possible explanation is that professional and community leadership are distinct concepts. Two of the three components—shows professionalism (4b) and supports professional growth (4c)—pertain to professional leadership, while the other component—maximizes parent and community involvement (4a)—pertains to community leadership. Empirically, excluding component 4a doubles the internal consistency of domain 4 for assistant principals from 0.20 to 0.41 (table D2 in appendix D). This evidence suggests that the types of external outreach measured by component 4a may not relate to the same underlying leadership concept as the more internally focused professional leadership measured by the domain's other components.

The low internal consistency of domain 4 for assistant principals also may reflect the possibility that the responsibilities of participating assistant principals did not include all the components in the domain, even though they were rated on those components. For example, some assistant principals could have little to no involvement with community outreach or teachers' professional development. To ensure that FFL scores reflect a coherent assessment of assistant principals' performance on the actual duties they are assigned, a supervisor may need to review an assistant principal's responsibilities before determining the components that factor into the domain scores, particularly in domain 4.

Across all components, 95 percent of principals and assistant principals were rated as either proficient or distinguished

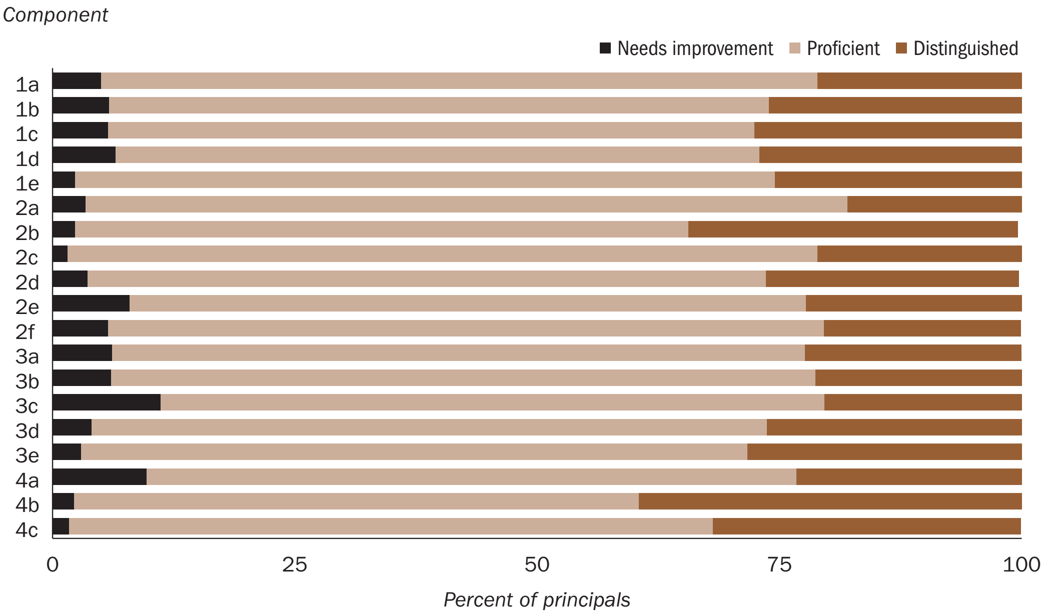
Variation in Framework for Leadership scores was very limited, with most scores in the top two of four performance categories

Score variation indicates whether the evaluation tool can differentiate levels of performance. Prior research has shown that principals differ considerably in their effectiveness in raising student achievement (Branch et al., 2012; Chiang et al., 2012; Coelli & Green, 2012; Dhuey & Smith, 2012a, 2012b). FFL scores should thus vary considerably as well.

Because component scores are inputs into domain scores, which are inputs into full FFL scores, the analysis begins by examining variation in component scores and then looks at variation in domain scores and full FFL scores. Two approaches were used to calculate domain scores. The first approach, used throughout this report, calculates each domain score as the unrounded, equal-weighted average of component scores for the domain (see box 1). To explore how score variation might differ under the Pennsylvania Department of Education's plan for supervisors to assign a whole-number domain score by judging the preponderance of evidence within the domain, the second approach rounds each domain score from the first approach to the nearest whole number. Under both approaches the full FFL score is the equal-weighted average of domain scores.

On every component, principals and assistant principals were most likely to be rated proficient or distinguished. On average across all components, 95 percent of principals and assistant principals were rated as either proficient or distinguished (figures 1 and 2; tables E1 and E2 in appendix E). The most common rating of performance on any FFL component was proficient (58–79 percent of principals and 64–91 percent of assistant principals), followed by distinguished (18–40 percent of principals and 5–36 percent of assistant principals). On average, supervisors assigned the needs improvement rating about

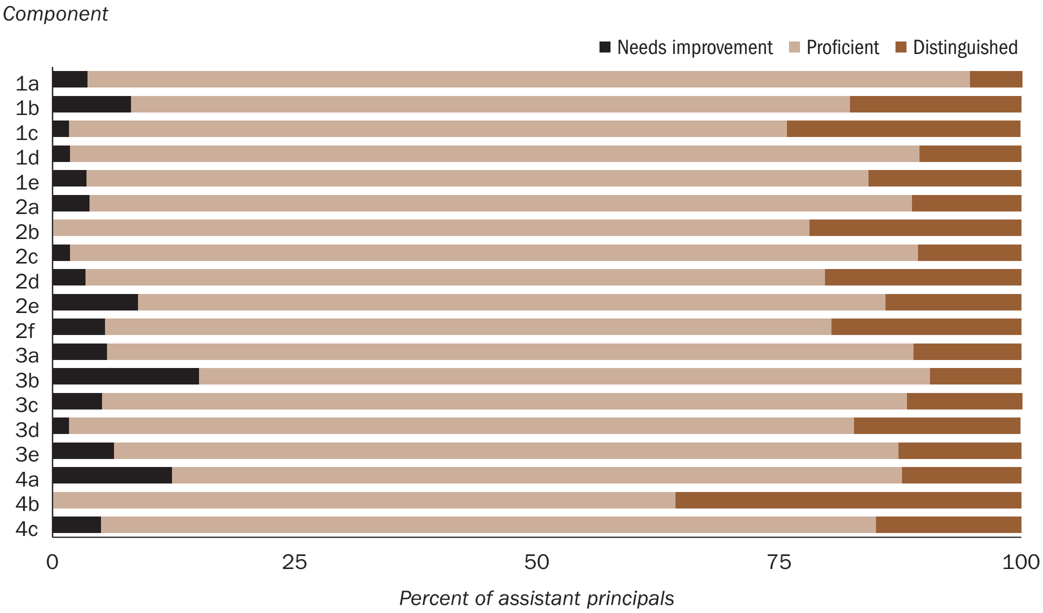
Figure 1. On every component of the Framework for Leadership, principals were most frequently rated as proficient or distinguished in the 2012/13 pilot year



Note: Only two principals received a failing score on a component. See table B1 in appendix B for definitions of components.

Source: Authors' calculations based on Framework for Leadership 2012/13 pilot evaluation scores provided by the Pennsylvania Department of Education.

Figure 2. On every component of the Framework for Leadership, assistant principals were most frequently rated as proficient or distinguished in the 2012/13 pilot year



Note: No assistant principals received a failing score on any component. See table B1 in appendix B for definitions of components.

Source: Authors' calculations based on the Framework for Leadership 2012/13 pilot evaluation scores provided by the Pennsylvania Department of Education.

5 percent of the time. Failing ratings were extremely rare: only two principals received a failing rating on a component, while no assistant principal received a failing rating on any of the 19 components. Variation in component scores was thus quite limited, with the vast majority of scores falling into the top two of four possible performance categories.

For the FFL to provide fair evaluations when supervisors and school leaders choose the components to be rated, as they did for the 2012/13 pilot evaluations, the difficulty of scoring well should be about the same for each FFL component. Component score distributions for the pilot evaluations did not differ substantially across components (see figures 1 and 2). For both groups of school leaders, average component scores were similar in magnitude even after isolating differences in average scores across components due solely to differences in the difficulty of components rather than to differences in the mix of school leaders evaluated on different components (see table E3 and accompanying text in appendix E). Together with the earlier finding that the internal consistency of most FFL domains is marginally acceptable or better, these findings imply that allowing school leaders and their supervisors to choose which components to include in the evaluation may not compromise the fairness of the FFL scores across school leaders.

If domain scores were determined by the preponderance of evidence, the Framework for Leadership could not make any distinctions in performance among a majority of school leaders

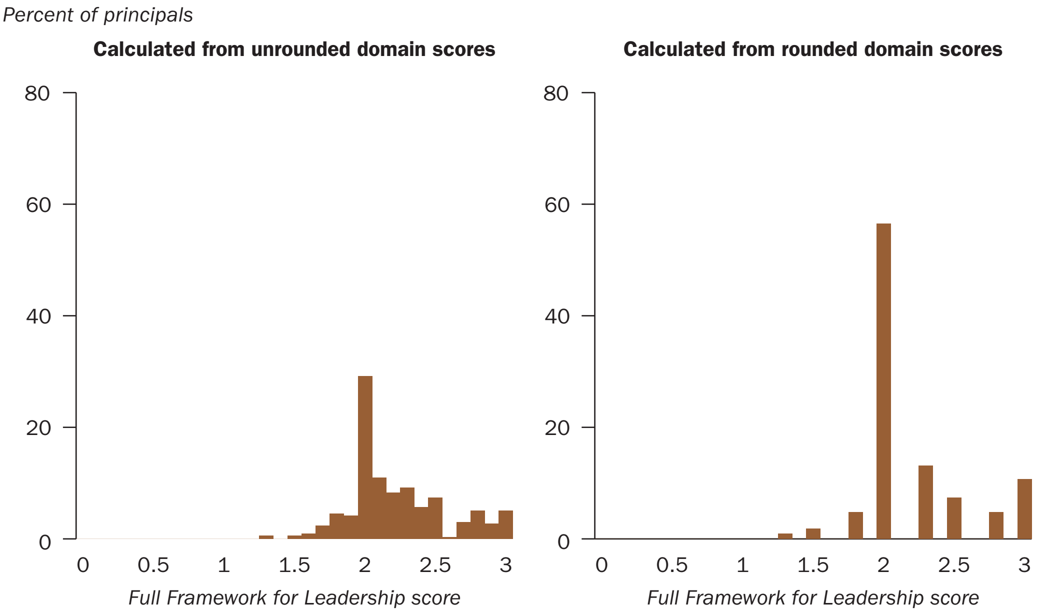
Scores for each domain and the full Framework for Leadership were concentrated at the top third of the scale. In view of the generally high component scores that school leaders received, the vast majority of leaders earned domain scores of 2.0 or above on the 0–3 point scale. In every domain the percentage of both principals and assistant principals scoring at least 2.0 exceeded 80 percent based on unrounded domain scores and 95 percent based on domain scores rounded to whole numbers (tables E4 and E5 in appendix E).

Likewise, full FFL scores for both principals and assistant principals were concentrated at the top third of the rating scale (figures 3 and 4). With full FFL scores calculated from unrounded domain scores, 83 percent of principals and 84 percent of assistant principals had a full FFL score of 2.0 or higher (see tables E4 and E5 in appendix E). With full FFL scores calculated from rounded domain scores, the corresponding percentages were 93 and 97 percent. The most common full FFL score was exactly 2.0 (25 percent of principals and 29 percent of assistant principals based on unrounded domain scores; 57 percent of principals and 73 percent of assistant principals based on rounded domain scores).

Using preponderance of evidence to determine domain scores would reduce score variation. Rounding domain scores to whole numbers—as would be the case if supervisors assigned domain scores by judging the preponderance of evidence—lowers the variation in full FFL scores compared with specifying domain scores to be unrounded averages of component scores. There were fewer distinct values for the full FFL scores when they were calculated from rounded rather than unrounded domain scores (see figures 3 and 4). Moreover, because most unrounded domain scores were within 0.5 point below or above 2.0, rounding those domain scores to 2 would eliminate all distinctions among school leaders in that range of scores. As a result, a majority of school leaders would earn a 2 on every domain and thus have the identical full FFL score of 2 (see the right panels of figures 3 and 4). In other words, if domain scores were determined by the preponderance of evidence, the FFL could not make any distinctions in performance among a majority of school leaders.

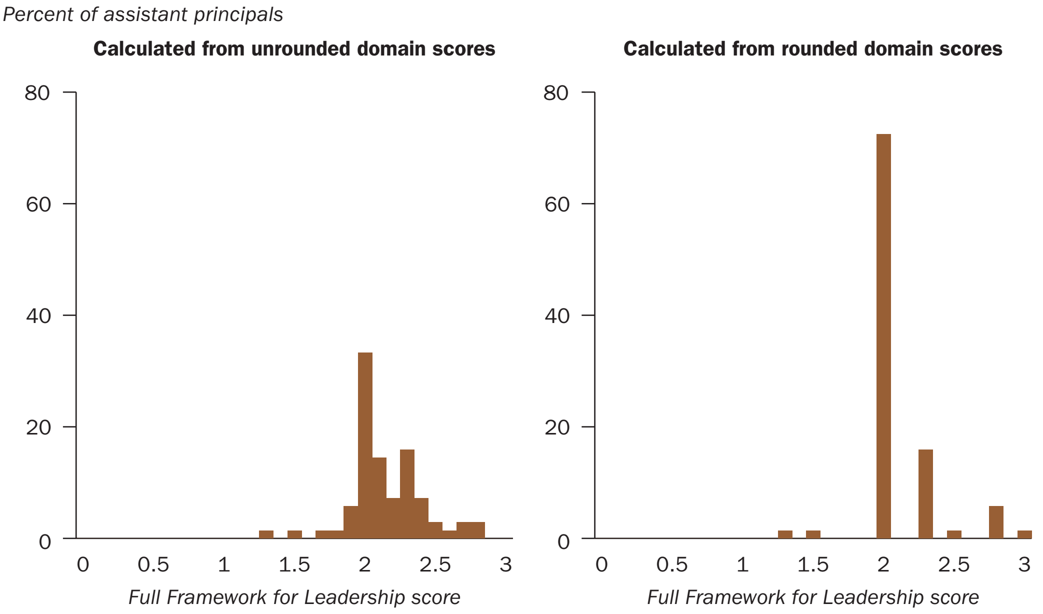
Although FFL scores varied somewhat, most school leaders in the 2012/13 pilot received high scores. The prevalence of high scores could mean that supervisors did not sufficiently

Figure 3. Full Framework for Leadership scores were concentrated at the top third of the scale among principals in the 2012/13 pilot year



Source: Authors' calculations based on Framework for Leadership 2012/13 pilot evaluation scores provided by the Pennsylvania Department of Education.

Figure 4. Full Framework for Leadership scores were concentrated at the top third of the scale among assistant principals in the 2012/13 pilot year



Source: Authors' calculations based on Framework for Leadership 2012/13 pilot evaluation scores provided by the Pennsylvania Department of Education.

differentiate between levels of performance or that highly effective leaders were most likely to participate in the pilot. One way to distinguish between these possibilities is to see whether FFL score patterns can be substantiated by other evidence.

Framework for Leadership scores did not correlate with estimates of school leaders' contributions to student achievement growth

If the FFL is working as intended, the FFL scores should be positively related to school leaders' contributions to student achievement growth. This is because the Pennsylvania Department of Education regards the leadership practices measured by the FFL as school leaders' key inputs into improving student achievement. The strength of this relationship was assessed by correlating school leaders' 2012/13 FFL scores with an objective measure of their contributions to student achievement growth in the same year. Because both the FFL scores and the objective measure to which they are compared are supposed to capture school leaders' effectiveness in the same school year (2012/13), this analysis provides an assessment of the FFL's concurrent validity.

The school leaders in the pilot received high Framework for Leadership scores even though their estimated contribution to student achievement growth was about average for the state

This study measures school leaders' contributions to student achievement growth using a value-added model. The effectiveness of the leaders' schools in 2012/13—captured by how much student achievement growth that year exceeded or fell below predictions—was the starting point for measuring leaders' contributions (see appendix F for details). The value-added measure was refined for recently hired leaders but not for longer-serving leaders, to account for differences in school effectiveness that resulted from actions taken by previous school leaders. Value-added estimates for recently hired leaders who began their current positions in 2008/09 or later were adjusted using data on school effectiveness prior to the leaders' arrival. These data were not available for longer-serving leaders who began their current positions before 2008/09, so school effectiveness in 2012/13 was used as a proxy for the leaders' own effectiveness. Because the resulting value-added estimates have greater validity for recently hired leaders than for longer-serving leaders, the relationships between value-added and FFL scores were estimated separately for these two groups.[4]

The FFL's concurrent validity could vary depending on whether components, domains, or the full FFL is considered. The domain and component scores with the largest positive associations with value-added could represent promising practices for the Pennsylvania Department of Education to target for professional development. Findings could also vary depending on whether estimates of leaders' value-added are based on student outcomes in all tested subjects combined or in particular subjects. Finally, findings could vary for principals and for assistant principals and by the grade span of the leaders' schools. Thus, the study estimated relationships for all these combinations.

The estimated value-added of school leaders in the 2012/13 pilot was not above the statewide average. As a group, the leaders in the pilot received high FFL scores even though their estimated value-added was about average for the state, suggesting that supervisors rated leniently. For the three key groups of leaders considered in the analysis—recently hired principals, longer-serving principals, and recently hired assistant principals—and across nearly all subjects, the estimated average value-added of pilot participants was statistically indistinguishable from the average value-added of all school leaders in the state (see table F7 and accompanying text in appendix F).

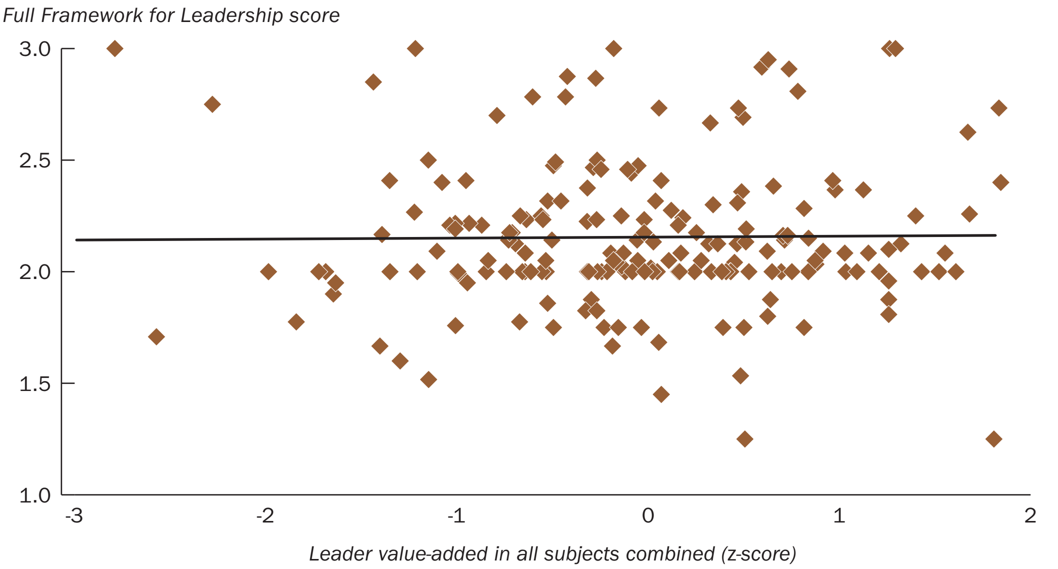
Differences in FFL scores among leaders could still indicate differences in value-added, giving rise to a relationship between these measures. This possibility is examined next.

Framework for Leadership scores were unrelated to school leaders' estimated value-added. School leaders' FFL scores in the 2012/13 pilot did not have a statistically significant relationship with estimated contributions to student achievement growth. For all three key groups of leaders—recently hired principals, longer-serving principals, and recently hired assistant principals—neither the full FFL scores nor any of the domain scores were associated with the school leaders' estimated value-added in all subjects combined (figure 5; see also tables G1, G2, and G3 in appendix G). For example, the almost perfectly horizontal line in figure 5 indicates that recently hired principals with greater contributions to achievement growth received FFL scores that, on average, were no better or worse than the FFL scores earned by recently hired principals whose estimated contributions were smaller. Findings were similar for longer-serving principals and recently hired assistant principals.

Neither the full Framework for Leadership scores nor any of the domain scores were associated with school leaders' estimated value-added

There were also no statistically significant relationships between full or domain-level FFL scores and school leaders' estimated value-added in particular subjects, including math, reading and writing, and science (see tables G1, G2, and G3 in appendix G). In addition, in nearly all cases, FFL component scores were unrelated to leaders' estimated value-added. Of the 57 estimated relationships between FFL component scores and leaders' value-added in all subjects combined, only one was statistically significant—a smaller number of significant estimates than would be expected based on pure chance (table G4 in appendix G). Finally, when principals were divided into three groups based on the grade span of the school they led—elementary, middle, and high school—there were no statistically significant relationships between school leaders' full or domain-level FFL scores and their estimated value-added in all subjects combined (table G5 in appendix G).[5]

Figure 5. There was no relationship between full Framework for Leadership scores and estimated value-added scores for recently hired principals



Note: Recently hired principals began their current positions in 2008/09 or later.

Source: Authors' calculations based on Framework for Leadership 2012/13 pilot evaluation data, student achievement and background data, and school leaders' job assignment data provided by the Pennsylvania Department of Education.

Implications and limitations of the study

This study, one of the few to document the internal consistency, score variation, and concurrent validity of a school leader evaluation tool intended for large-scale implementation, found that the FFL exhibited a mix of strengths and weaknesses in the 2012/13 pilot.

A key strength of the FFL is its internal consistency. School leaders identified as effective or ineffective on one domain of the FFL tended to be identified in a similar way on the other domains. This internal consistency could be explained by a "halo effect," with supervisors assigning the same high score to every component if they think a subordinate's performance is good. With no evidence either supporting or ruling out a halo effect, this study assumes the benign explanation for the FFL's high internal consistency: different domains of the FFL capture a common underlying definition of effective school leadership.

However, the FFL in its pilot phase is not yet meeting its objective of differentiating principals who make greater or smaller contributions to student achievement growth. Most school leaders scored in the upper third of the rating scale, which suggests a tendency for supervisors to rate their school leaders too leniently. Moreover, when FFL scores were calculated from domain scores rounded to whole numbers, the variation in FFL scores was further reduced, suggesting that using the preponderance of evidence to determine domain scores would further decrease the FFL's ability to differentiate levels of performance. In contrast, prior research that examined variation in value-added across leaders found considerable differences in school leaders' contributions to achievement growth. Therefore, the concentration of FFL scores in a narrow range of the scale is one indication that the scores in the pilot may not have strongly reflected school leaders' contributions to achievement growth. This implication is substantiated by direct evidence: principals with greater estimated contributions to student achievement growth did not, on average, score higher on the FFL than principals with smaller estimated contributions.

This study, one of the few to document the internal consistency, score variation, and concurrent validity of a school leader evaluation tool intended for large-scale implementation, found that the FFL exhibited a mix of strengths and weaknesses in the 2012/13 pilot

This finding does not necessarily imply that FFL scores are less valid indicators of school leader effectiveness than scores from other evaluation tools. Examining the validity of the FFL scores on the basis of their relationship with estimates of school leaders' value-added sets a very high bar for the FFL. Almost no studies have documented a relationship of evaluation tools with school leaders' value-added. Two exceptions focused on a small number of district-specific evaluation instruments and did not find any robust evidence of a relationship between these instruments and principals' value-added (Milanowski & Kimball, 2012; Grissom, Kalogrides, & Loeb, 2012). And no studies have analyzed how school leader evaluation tools relate to the value-added of assistant principals. Thus, it is unknown whether other evaluation tools would be any more indicative of school leaders' value-added than the FFL.

Nevertheless, the absence of a relationship between FFL scores and estimates of school leaders' value-added suggests that more evidence is needed on the validity of using FFL scores to identify effective and ineffective school leaders. Specifically, if average FFL scores continue to be high, it will be important to determine whether other evidence can support the conclusion that the evaluated leaders, on the whole, exhibit good leadership. Likewise, it will be important to learn whether differences in FFL scores provide meaningful information about performance differences that are corroborated by other evidence.

Limitations of the study

Interpretation of this study's findings should consider several limitations.

Focus on test-based achievement outcomes. In this study all measures of school leaders' contributions to achievement growth were based on student outcome measures from state tests. This study does not examine whether FFL scores reflect school leaders' contributions to student outcomes that are not reflected in state tests—such as creativity and character. In subsequent years, a follow-up study will use student enrollment data in a value-added framework to measure high school leaders' contributions to their students' enrollment persistence (avoidance of dropout). This will permit additional analyses of the extent to which FFL scores reflect those contributions.

Lack of consensus on how to measure school leaders' value-added on a large scale. This study developed a new method for measuring school leaders' value-added (see appendix F). Previous studies that measured principals' value-added used flawed methods that attribute the effectiveness of entire schools to the effectiveness of the principal alone or that typically compare principals on their effectiveness only if they have led the same school during the period under study (Branch et al., 2012; Chiang et al., 2012; Coelli & Green, 2012; Dhuey & Smith, 2012a, 2012b; Grissom et al., 2012; and Lipscomb, Chiang, & Gill, 2012). Although this study developed a method for comparing effectiveness among a larger group of school leaders, there is no clear consensus on the most theoretically satisfying and practically realistic method for large-scale comparisons of school leaders' value-added.

The absence of a relationship between Framework for Leadership (FFL) scores and estimates of school leaders' value-added suggests that more evidence is needed on the validity of using FFL scores to identify effective and ineffective school leaders

Flawed measure of the value-added of longer-serving school leaders. As discussed in appendix F, this study's most valid measure of school leaders' value-added relied on accounting for the effectiveness of a leader's school before the leader arrived. However, the study lacked data on school performance before the arrival of longer-serving school leaders—those who began their current positions before 2008/09—and therefore could not control for the lingering effects of these leaders' predecessors. This decreased the validity of the value-added measures for these school leaders, so the study's analysis of the relationships between FFL scores and value-added is less valid for these leaders than for recently hired leaders.

Limited sample size. The sample size for the 2012/13 pilot did not permit very precise estimates of the relationship between school leaders' FFL scores and their value-added. For the estimates to have been reliably statistically significant, the FFL would need to reflect school leaders' value-added at least as strongly as the Framework for Teaching (the classroom observation tool used in Pennsylvania; see appendix G) reflects teachers' value-added. Given that the Framework for Teaching is a well established tool while the FFL is new, it is possible that some FFL components may have a real but smaller relationship with value-added that was too small for this study to detect. The sample size for assistant principals was particularly small, and findings for assistant principals could change when a larger sample of assistant principals is available for analysis in the next pilot phase.

Suggestions for improving FFL evaluations and gathering more evidence on its validity

Central questions that arise from this study's findings are whether differences in FFL scores among school leaders offer meaningful information about differences in leaders' performance, whether supervisors are too lenient when assigning FFL scores, and how the internal consistency of domain 4 (professional and community leadership) can be improved. The Pennsylvania Department of Education could provide supervisors with more guidance on assigning scores, gather additional evidence that can corroborate or refute conclusions about performance based on the FFL scores, refine domain 4, and collect pilot ratings from all participating school leaders.

Provide more guidance to supervisors on how to assign scores for each component. More specific guidance would help supervisors determine whether they are rating appropriately. Supervising administrators in the 2012/13 pilot participated in a one-day training session to familiarize themselves with the FFL, but they received only general guidance on how to assign scores. They were given definitions of the four performance categories (distinguished, proficient, needs improvement, and failing) tailored to each component, as well as lists of the types of evidence that school leaders could submit to inform their evaluations. The Pennsylvania Department of Education could provide illustrative, concrete examples of the quantity and quality of evidence that would merit each possible score for every FFL component, which would enable supervisors to refer to those examples when assessing the evidence presented by school leaders. To the extent that these examples set a higher standard for scoring well than the personal standards that supervisors used in 2012/13, fewer FFL scores will be concentrated in upper parts of the rating scale, leading to greater score variation.

The Pennsylvania Department of Education could provide examples of the quantity and quality of evidence that would merit each possible score for every FFL component, which would enable supervisors to refer to those examples when assessing the evidence presented by school leaders

Obtain ratings of school leaders by other stakeholders to check the validity of scores assigned by the supervisors. Ratings of school leaders by knowledgeable individuals other than supervisors can provide general statistical information on the validity of the supervisors' conclusions, even if the ratings do not factor officially into evaluations of the school leaders. In particular, asking teachers to rate their school leaders anonymously using the FFL could yield informative results. This is analogous to using student surveys as part of teacher evaluations, a practice found to improve the reliability and validity of teacher effectiveness measures (Kane & Staiger, 2012). It is also consistent with the "360" evaluations commonly used in the corporate world. While this approach might necessitate selecting components of the FFL that teachers are equipped to assess, ratings by teachers may be less susceptible to excessive leniency due to their anonymity. And because the ratings would include the perspectives of many observers, they are likely to have reasonable levels of reliability.

Gathering additional evidence from ratings by teachers would enable the Pennsylvania Department of Education to compare average scores based on teachers' ratings with average scores based on supervisors' ratings to assess whether supervisors are being too lenient or too strict. And it would enable the Pennsylvania Department of Education to assess the FFL's convergent validity—the extent to which differences in school leaders' scores based on one approach (ratings by supervisors) are reflected in corresponding differences based on another approach (ratings by teachers). Taken together, this evidence would be valuable in establishing the FFL's validity.

Improve the internal consistency of domain 4. The internal consistency findings for domain 4 (professional and community leadership) suggest that the domain may need further refinement, particularly for assistant principals. The internal consistency of domain 4 could be improved by adding more components to the domain that would apply to both professional leadership and community leadership or by splitting domain 4 into two domains. A supervisor may need to carefully review an assistant principal's responsibilities before determining the components that ought to factor into the school leader's domain 4 scores.

Ensure that pilot ratings are collected from all participating school leaders in the 2013/14 pilot year. As noted earlier, the small number of school leaders (particularly assistant principals) who submitted rating data in the 2012/13 pilot made it challenging to detect relationships between FFL scores and school leaders' value-added, even if true relationships exist. For the 2013/14 pilot year the Pennsylvania Department of Education projected that approximately 1,200 principals and 500 assistant principals would participate in pilot evaluations. The Pennsylvania Department of Education should ensure that evaluation data are submitted to the study for all pilot participants. Analyses of recently hired and longer-serving principals and recently hired assistant principals should thus be sufficiently precise to detect any relationships that might be considered meaningful in magnitude (see appendix G for details).

Appendix A. Prior research on measuring principal effectiveness

The reliability and validity of most evaluation tools for rating school leaders are unknown. A review of 65 principal evaluation tools used by districts and states receiving Wallace Foundation grants revealed that 63 of those tools had no documentation of their reliability or validity (Goldring et al., 2009). A keyword search in Google Scholar conducted by Condon and Clifford (2012) found only eight evaluation tools with any information on reliability or validity. With the few exceptions described below, the available statistical information on these evaluation tools typically consists only of measures of reliability and a very limited form of validity (construct validity), assessing whether conceptual groupings of components in those tools can be empirically verified by confirmatory factor analysis or other methods.

Only a few studies have developed and analyzed methods for estimating principals' contributions to student achievement growth (Branch et al., 2012; Chiang et al., 2012; Coelli & Green, 2012; Dhuey & Smith, 2012a, 2012b; Grissom et al., 2012; Lipscomb et al., 2012). These methods are based on value-added models, which are analytic models that control for students' prior achievement and demographic characteristics when comparing student growth across teachers, schools, or school leaders. The resulting measures of effectiveness are known as value-added measures. A key observation from this research is that a principal's value-added is not the same as the value-added of the school that he or she leads, because the school's value-added may also reflect other school-specific factors beyond the principal's control (Chiang et al., 2012). For example, the composition of a school's teaching staff is likely to influence the school's value-added, and a school may inherently find it relatively easy or difficult to attract good teachers due, for instance, to neighborhood characteristics.

One common method of distinguishing principals' value-added from the influence of other school-specific factors is to compare the same school's performance under two different principals. The more effective principal is the one under whom the school fared better. Because student outcomes under both principals are for the same school, this method controls for all school-specific factors that do not change over time. However, this method is unsuitable for a large-scale evaluation system because it can be applied only to schools with principal turnover during the period considered and, in most cases, can compare each principal only to other principals who have served the same school (Lipscomb et al., 2012). For this reason, this study developed a different method for estimating principals' contributions to student achievement growth (see appendix F).

Despite the recent methodological developments in value-added estimation, there is no consistent evidence that any principal evaluation tool currently in use produces scores that reflect the principals' value-added. For most principal evaluation tools, no empirical evidence is available about relationships between scores and student achievement growth. For example, none of the tools examined by Goldring et al. (2009) and Condon and Clifford (2012) has documentation of relationships with student achievement growth. To date, only two studies spanning three districts have examined the relationship between principal evaluation tools and value-added. In one such study based on two anonymous, medium-size districts, principals' scores were generally uncorrelated with school value-added in reading and math, although in math the correlations were statistically significant in a minority of the analysis samples considered (Milanowski & Kimball, 2012). In Miami-Dade County

Public Schools, principals' scores were positively associated with the value-added of their schools but did not have a robust association with value-added measures that specifically distinguished principals' contributions from the influence of other school-specific factors (Grissom et al., 2012).

Developers of some principal evaluation tools have assessed their validity through approaches other than examining relationships with principal value-added. For example, one recently developed tool, the Vanderbilt Assessment of Leadership in Education, has been the subject of several validity studies (Porter et al., 2008). An examination of the tool's convergent validity—the extent to which different measurement methods using the same tool produced similar scores—found that ratings of the same principal by different stakeholders (teachers, supervisors, and the principals themselves) had positive correlations in the range of 0.13 to 0.27 (Porter et al., 2010). In an analysis of the tool's concurrent validity—its relationship with another measure of the same concepts—teachers' ratings of their principals using the Vanderbilt Assessment of Leadership in Education had a positive correlation of 0.7 with ratings using a different tool, the Principal Instructional Management Rating Scale (Goldring, Cravens, Murphy, Porter, & Elliot, 2012). A "known group" validity study found that principals who were subjectively identified by superintendents as being in the top 20 percent of principals in their district scored higher on the Vanderbilt Assessment of Leadership in Education, based on principals' self-ratings and teachers' ratings, than those identified as being in the bottom 20 percent (Covay et al., 2013).

Appendix B. Structure of the Framework for Leadership

The Framework for Leadership (FFL) specifies 19 leadership practices, known as components, on which each school leader is rated by an administrator who has supervisory authority over the school leader (table B1). A school leader can receive a score of distinguished (3 points), proficient (2 points), needs improvement (1 point), or failing (0 points) on each component. School leaders also receive a summary score (with the same possible 3, 2, 1, or 0 points) for each domain, based on the preponderance of evidence from the component scores. The ratings supervisors assign are based on direct observation and on evidence submitted by the school leaders.

Table B1. Components of the Framework for Leadership, by domain

Name of component	Description of component
1: Strategic/cultural leadership	
1a. Creates an organizational vision, mission, and strategic goals	The school leader plans strategically and creates an organizational vision, mission, and goals around personalized student success that are aligned to local education agency goals.
1b. Uses data for informed decisionmaking	The school leader analyzes and uses multiple data sources to drive effective decisionmaking.
1c. Builds a collaborative and empowering work environment	The school leader develops a culture of collaboration, distributive leadership, and continuous improvement conducive to student learning and professional growth. The school leader empowers staff in the development and successful implementation of initiatives that better serve students, staff, and the school.
1d. Leads change efforts for continuous improvement	The school leader systematically guides staff through the change process to positively impact the culture and performance of the school.
1e. Celebrates accomplishments and acknowledges failures	The school leader utilizes lessons from accomplishments and failures to positively impact the culture and performance of the school.
2: Systems leadership	
2a. Leverages human and financial resources	The school leader establishes systems for marshaling all available resources to better serve students, staff, and the school.
2b. Ensures school safety	The school leader ensures the development and implementation of a comprehensive safe schools plan that includes prevention, intervention, crisis response, and recovery.
2c. Complies with federal, state, and local education agency mandates	The school leader designs protocols and processes to comply with federal, state, and local education agency mandates.
2d. Establishes and implements expectations for students and staff	The school leader establishes and implements clear expectations, structures, rules, and procedures for students and staff.
2e. Communicates effectively and strategically	The school leader strategically designs and utilizes various forms of formal and informal communication with all staff and stakeholders.
2f. Manages conflict constructively	The leader effectively and efficiently manages the complexity of human interactions and relationships, including those among and between parents/guardians, students, and staff.
3: Leadership for learning	
3a. Leads school improvement initiatives	The school leader develops, monitors, and evaluates a School Improvement Plan that provides the structure for the vision, goals, and changes necessary for improved student achievement.
3b. Aligns curricula, instruction, and assessments	The school leader ensures that the adopted curricula, instructional practices, and associated assessments are implemented within a Standards Aligned System. Data are used to drive refinements to the system.

(continued)

Table B1. Components of the Framework for Leadership, by domain *(continued)*

Name of component	Description of component
3c. Implements high-quality instruction	The school leader monitors progress of teachers and staff. In addition, the school leader conducts formative and summative assessments in measuring teacher effectiveness to ensure that rigorous, relevant, and appropriate instruction and learning experiences are delivered to and for all students.
3d. Sets high expectations for all students	The school leader holds all staff accountable for setting and achieving rigorous performance goals for all students.
3e. Maximizes instructional time	The school leader creates processes that protect teachers from disruption of instructional and preparation time.
4: Professional and community leadership	
4a. Maximizes parent and community involvement and outreach	The school leader designs structures and processes that result in parent and community engagement, support, and ownership for the school.
4b. Shows professionalism	The leader operates in a fair and equitable manner with personal and professional integrity.
4c. Supports professional growth	The school leader supports continuous professional growth of self and others through practice and inquiry.

Source: Pennsylvania Department of Education.

Appendix C. Data used in the study

The study used data on Framework for Leadership (FFL) scores and other individual-level administrative data on students and school leaders in Pennsylvania. This appendix provides details on these data sources.

The 2012/13 pilot year: Participants, evaluation procedures, and available data

Participants. All the FFL scores used by this report came from the 2012/13 FFL pilot year. Understanding the criteria for participation in the 2012/13 pilot year and the characteristics of the participants can shed light on the types of schools and school leaders to whom the findings pertain.

The school leaders whose FFL scores were used in the analysis came from 344 schools spread across 146 local education agencies (table C1). The study's analyses included 405 school leaders—336 principals and 69 assistant principals—with FFL scores from the pilot year. Collectively, these leaders were rated by 171 supervisors.

Local education agencies and schools that participated in the 2012/13 pilot year did so for one of three reasons. First, local education agencies receiving Race to the Top funds were required to select at least one school to participate. Second, schools receiving School Improvement Grants to implement a transformation model of improvement were required to participate. Third, local education agencies could voluntarily select schools to participate. The large majority of local education agencies in the study (116 of 146) were required to participate because they received Race to the Top funds (table C2). Most of the principals (281 of 336) and assistant principals (63 of 69) in the study were leaders in these 116 local education agencies.

Characteristics of students enrolled in schools that did and did not participate in the 2012/13 pilot year are shown in table C3; characteristics of participating and nonparticipating school leaders are shown in table C4.

Evaluation procedures. One supervising administrator evaluated each school leader in the pilot. Superintendents and assistant superintendents constituted the majority of supervisors who rated principals (82 percent) and assistant principals (66 percent; figure C1). One-fourth of the supervisors who rated assistant principals were the principals to whom the assistant principals were accountable.

Table C1. Number of participants in the 2012/13 Framework for Leadership pilot year

Type of participant	Number
Local education agencies (districts, charter schools, technical centers)	146
Schools	344
School leaders who received ratings	405
Principals	336
Assistant principals	69
Supervisors who assigned ratings	171

Source: Authors' calculations based on Framework for Leadership 2012/13 pilot evaluation scores provided by the Pennsylvania Department of Education.

Table C2. Reasons for the participation of local education agencies in the Framework for Leadership 2012/13 pilot year

Reason for participation of local education agency	Number of local education agencies	Number of principals	Number of assistant principals
Receives Race to the Top Funds (and no other reason)	104	243	58
Receives Race to the Top Funds and has school receiving School Improvement Grant funds for transformation	12	38	5
Has school receiving School Improvement Grant funds for transformation (and no other reason)	5	13	2
Volunteer	23	40	4
Reason not recorded	2	2	0

Source: Authors' calculations based on Framework for Leadership 2012/13 pilot evaluation scores provided by the Pennsylvania Department of Education.

Table C3. Characteristics of students in Pennsylvania in 2012/13, by whether their school participated in the Framework for Leadership 2012/13 pilot year (percent unless otherwise indicated)

	Grades 4–5		Grades 6–8		Grades 9–12	
Student characteristic	School did not participate	School participated	School did not participate	School participated	School did not participate	School participated
Number of students	223,386	24,719	336,803	49,750	300,847	49,989
Baseline math score (average *z*-score)	0.01	0.00	0.02	0.00	0.09	0.01
Baseline reading score (average *z*-score)	0.01	–0.01	0.02	–0.02	0.08	–0.01
Receives free lunch	39.1	38.8	36.4	36.3	29.7	34.7
Receives reduced-price lunch	5.2	6.1	5.4	6.3	5.3	5.5
English language learner student	2.4	1.4	2.2	1.5	1.3	1.9
Any disability	17.2	17.3	16.6	16.6	14.0	13.4
Moved schools during school year	3.8	3.6	4.4	4.1	11.4	9.5
Grade repeater	0.2	0.3	0.7	0.7	4.0	4.3
Over age for grade	0.2	0.2	0.3	0.3	0.9	1.0
Age (average years)	10.1	10.1	12.6	12.6	15.7	15.7
Female	48.9	49.2	48.8	48.8	49.3	49.4
Race/ethnicity						
Asian or Pacific Islander	3.7	2.0	3.3	2.0	3.1	2.1
Black, non-Hispanic	14.6	15.0	14.7	10.8	13.2	11.4
Hispanic	9.2	5.3	8.1	6.4	6.4	7.6
White, non-Hispanic	69.6	75.0	70.4	77.0	73.4	75.0
Other race/ethnicity	2.1	2.0	1.1	0.9	0.4	0.6

Note: Statistics in the table are based only on students who were included in at least one value-added model described in appendix F. For students in grades 4–8, baseline scores come from the previous year; for students in grades 9–12, baseline scores come from grade 8.

Source: Authors' calculations based on student achievement and background data provided by the Pennsylvania Department of Education.

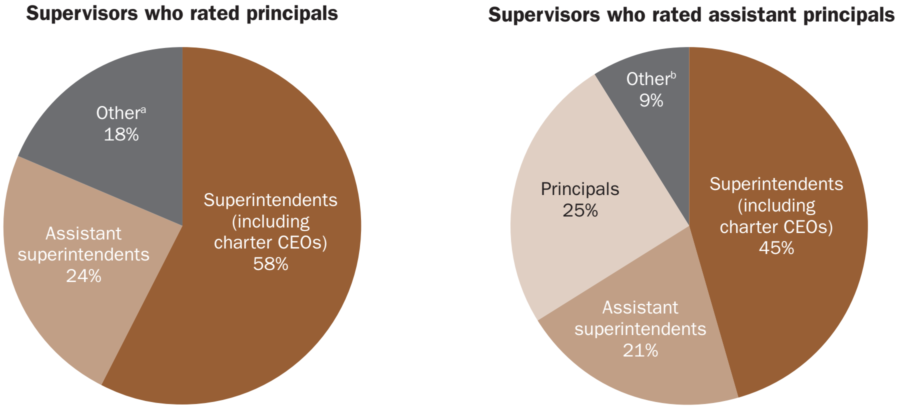
Table C4. Characteristics of school leaders in Pennsylvania, by whether they participated in the Framework for Leadership 2012/13 pilot year (percent unless otherwise indicated)

Characteristic	Principals who did not participate	Principals who participated	Assistant principals who did not participate	Assistant principals who participated
Highest degree attained				
Bachelor's	15.4	11.8	14.0	9.1
Master's	74.2	79.0	82.0	84.4
Doctorate	9.7	8.6	2.9	2.6
Total experience in PK–12 education (average years)	19.1	17.1	15.1	13.8
Race and ethnicity				
Black, non-Hispanic	10.8	8.0	13.5	7.8
White, non-Hispanic	86.6	90.2	82.3	88.3
Other	1.6	1.2	2.8	0.0
Gender				
Female	45.4	35.5	42.9	23.4
Male	53.2	63.3	55.1	72.7

PK–12 is prekindergarten to grade 12.

Source: Authors' calculations based on job assignment and background data on school leaders provided by the Pennsylvania Department of Education.

Figure C1. Most supervisors in the Framework for Leadership 2012/13 pilot year were superintendents or assistant superintendents



a. Includes other principals, directors of vocational education, supervisors of curriculum and instruction, supervisors of elementary education, and supervisors of secondary education.

b. Includes supervisors of curriculum and instruction.

Source: Authors' calculations based on Framework for Leadership 2012/13 pilot evaluation scores provided by the Pennsylvania Department of Education.

The state's intermediate units (regional agencies that provide instructional and operational services to groups of school districts) were responsible for training supervisors in using the FFL. Training in the 2012/13 pilot year occurred in two stages. First, staff from the Pennsylvania Department of Education conducted a two-day "train-the-trainer" session for intermediate unit leaders to familiarize them with the FFL and guide them in facilitating training activities for supervisors. The train-the-trainer session covered general topics, such as:

- The background and rationale for the FFL.
- The state of the research on principal effectiveness.
- The specific domains measured by the FFL.
- The definitions of the four performance categories (distinguished, proficient, needs improvement, and failing) tailored to each component.
- The types of evidence that school leaders might submit in each domain.
- Ways of integrating the FFL into districts' systems for school leader evaluation.

Next, intermediate unit leaders held one-day training sessions in their jurisdictions for the supervising administrators who would be rating school leaders. These one-day sessions covered topics similar to those in the train-the-trainer session. Neither type of training session discussed concrete examples of the quantity and quality of evidence that would merit each possible score for every FFL component.

Participants in the pilot had some discretion over which FFL components would be included in the pilot evaluations. According to guidance from the Pennsylvania Department of Education, each pilot evaluation was supposed to include at least three components spread across at least two domains, representing a mix of the school leaders' strengths and weaknesses. School leaders and their supervisors were instructed to meet at the beginning of the school year to select components, devise goals for each component, and identify types of evidence that school leaders could submit for each component. They were also instructed to hold a midyear meeting to discuss progress toward the goals and an end-of-year meeting to review all evidence, culminating in final scores assigned by the supervisor at the end of the school year.

Available data. This study relies on FFL scores submitted by local education agencies to the Pennsylvania Training and Technical Assistance Network, an agency within the Pennsylvania Department of Education. Despite the discretion that school leaders and supervisors had in selecting components for the pilot evaluations, most of the pilot evaluations in the data included a component from every domain. The 405 school leaders in the analysis (see table C1) were evaluated on at least one component from every domain; they constitute 94 percent of an original group of 430 principals and assistant principals who had a score from any component. The 405 school leaders in the analysis were typically evaluated on most of the components; their pilot evaluations used an average of 16 out of 19 components, and 72 percent of the evaluations used all components.

Although actual FFL evaluations starting in 2014/15 will require supervisors to assign a domain score based on the preponderance of evidence within a domain, supervisors assigned only component scores in the 2012/13 pilot evaluations. For the analysis, the study computed a school leader's domain score as the equal-weighted average of scores from the components in the domain on which a school leader was evaluated. The Pennsylvania Department of Education regards the four domains as separate, equally weighted elements of a school leader's annual evaluation rating. The study's analyses of the full FFL required constructing a full FFL score, which the study defined as the equal-weighted average of the four domain scores.

Other administrative data on students and school leaders

Data on student achievement scores and background characteristics and school leaders' job assignments were necessary for estimating school leaders' contributions to student achievement growth. All of these data came from databases maintained by agencies at the Pennsylvania Department of Education.

The Pennsylvania Department of Education's Bureau of Assessment and Accountability provided the achievement scores of all students in the state who were administered state assessments from 2006/07 to 2012/13. The data covered the state's end-of-grade assessments, called the Pennsylvania System of School Assessment, which were administered in reading and math in grades 3–8 and grade 11; science in grades 4, 8, and 11; and writing in grades 5, 8, and 11. The data included modified Pennsylvania System of School Assessment tests administered to students with disabilities who were eligible for those assessments based on their individualized education program. The data also covered the state's end-of-course assessments, called the Keystone Exams, which were administered statewide for the first time in 2012/13, replacing the grade 11 Pennsylvania System of School Assessment tests. Keystone Exams were administered in algebra I, biology, and literature.

All other administrative data on students and school leaders came from the state's longitudinal data system, known as the Pennsylvania Information Management System, maintained by the Pennsylvania Department of Education. The data covered all students who were enrolled in the state's public schools and all principals and assistant principals who worked in those schools at any time from 2007/08 to 2012/13, and every student and educator in the data was assigned a unique identification number that was consistent across years. For each student in each year, the data indicated the schools in which the student was enrolled and information on the student's gender, age, race/ethnicity, free and reduced-price lunch status, English language learner status, and disability status. Data on principals and assistant principals indicated the schools to which they were assigned and information on their gender, education degrees, race/ethnicity, and total work experience in PK–12 education.

Appendix D. Technical details and supplementary findings on the internal consistency of the Framework for Leadership

This appendix provides technical details on how Cronbach's alpha (α) was calculated for the Framework for Leadership (FFL) and gives supplementary findings on internal consistency when particular domains or components were excluded.

Calculating Cronbach's alpha for the Framework for Leadership

The general formula for Cronbach's α to assess the internal consistency of a scale with k items is (Cronbach, 1951):

$$\text{(D1)} \qquad \alpha = \frac{k\bar{c}}{\bar{v} + (k-1)\bar{c}} \, ,$$

where $\bar{c}$ is the average covariance of item scores for all pairs of items and $\bar{v}$ is the average variance of item scores for all items.

Cronbach's α for the full FFL was obtained by treating the FFL as a scale with four items representing the four domain scores. The domain scores were calculated as equal-weighted averages among the components that were rated in each domain (regardless of which sets were rated for each school leader), because actual domain scores were not given in the pilot evaluation data for 2012/13. In actual evaluations, the Pennsylvania Department of Education plans to instruct supervisors to use the preponderance of evidence from the components in each domain to determine the domain scores.

Cronbach's α for a specific domain was obtained by treating the components within the domain as the items in applying equation D1. For each domain, the calculation is based on school leaders with scores on all components in the domain because the calculation of Cronbach's α relies on having complete data.

Supplementary findings on the internal consistency of the Framework for Leadership

Calculating α when particular domains or components are excluded from an index can provide supplementary information about the usefulness of parts of the index. If the resulting α values are appreciably lower than the α for the full index, the excluded piece is contributing positively to internal consistency. If the resulting α values are appreciably higher than the α for the full index, the excluded piece is contributing negatively to internal consistency. The α values obtained by excluding particular domains and components are provided in tables D1 and D2.

Table D1. Cronbach's alpha values for the full Framework for Leadership scores in the 2012/13 pilot year when particular domains are excluded

Portion of the Framework for Leadership used in calculating α	Cronbach's α	
	Principals	Assistant principals
Full Framework for Leadership with all four domains	0.88	0.85
Framework for Leadership, excluding:		
Domain 1: Strategic/cultural leadership	0.83	0.76
Domain 2: Systems leadership	0.85	0.87
Domain 3: Leadership for learning	0.85	0.81
Domain 4: Professional and community leadership	0.86	0.78

Source: Authors' calculations based on Framework for Leadership 2012/13 pilot evaluation scores provided by the Pennsylvania Department of Education.

Table D2. Cronbach's alpha values for Framework for Leadership domains in the 2012/13 pilot year when particular components are excluded

Portion of the Framework for Leadership used in calculating α	Cronbach's α	
	Principals	Assistant principals
Domain 1: Strategic/cultural leadership, excluding:		
No components	0.79	0.62
1a: Strategic goals	0.74	0.57
1b: Data for decisionmaking	0.75	0.55
1c: Empowering work environment	0.75	0.51
1d: Continuous improvement	0.72	0.55
1e: Lessons from accomplishments and failures	0.77	0.65
Domain 2: Systems leadership, excluding:		
No components	0.78	0.67
2a: Leverages resources	0.75	0.68
2b: School safety	0.76	0.65
2c: Complies with mandates	0.75	0.63
2d: Clear expectations for students and staff	0.74	0.53
2e: Communicates effectively	0.76	0.63
2f: Manages conflict	0.75	0.60
Domain 3: Leadership for learning, excluding:		
No components	0.82	0.65
3a: School improvement initiatives	0.79	0.55
3b: Aligns curricula and instruction	0.78	0.63
3c: High-quality instruction	0.77	0.53
3d: High expectations for students	0.77	0.58
3e: Maximizes instructional time	0.81	0.67
Domain 4: Professional and community leadership, excluding:		
No components	0.68	0.20
4a: Parent and community involvement	0.67	0.41
4b: Professionalism	0.50	0.04
4c: Supports professional growth	0.59	0.00

Source: Authors' calculations based on Framework for Leadership 2012/13 pilot evaluation scores provided by the Pennsylvania Department of Education.

Appendix E. Technical details and supplementary findings on variation in Framework for Leadership scores

This appendix provides detailed tabulations of the distribution of Framework for Leadership (FFL) scores. It also describes the methods used to compare average scores across components in a manner that adjusts for differences in the school leaders who were rated on different components.

Detailed tabulations of component score distributions

On every FFL component, the large majority of school leaders earned a score of either proficient or distinguished (see figures 1 and 2 in the main report). For each of the 19 components, detailed tabulations of the percentages of principals (table E1) and assistant principals (table E2) earning each of the four possible scores confirm this finding.

Formal analysis of component difficulty

When school leaders and their supervisors choose which components to use in an evaluation, school leaders may have an incentive to choose components that are substantially easier to score well in. Therefore, it is important to assess the difficulty of each component.

A component's difficulty can be reflected in school leaders' average score on the component. Lower average scores suggest greater difficulty. Average scores differed little across components, ranging from 2.1 to 2.4 for principals (see table E1) and from 1.9 to 2.4 for assistant principals (see table E2). These average scores constitute the first piece of evidence that the FFL components are similar in difficulty.

However, the average score on a component may also reflect the quality of school leaders who chose to be evaluated on the component. As discussed in appendix C, school leaders and their supervisors could choose which components to use in the pilot evaluations. To the extent that more (or less) effective school leaders chose to be rated on a component, average scores on the component will tend to be higher (or lower), regardless of the component's difficulty.

Further analytic steps were taken to isolate differences in average scores across components due solely to differences in the difficulty of components rather than to differences in the mix of school leaders evaluated on different components. These steps adjusted the differences in average scores across components to account for differences in the school leaders who were evaluated on those components. The data from all components and school leaders were pooled together into a common sample, separately for principals and assistant principals. For the numeric score on component *c* earned by school leader *i*, the following regression was estimated:

$$y_{ci} = \alpha_c + \theta_i + \in_{ci}, \qquad \text{(E1)}$$

where α_c is a fixed effect for component *c*, θ_i is a fixed effect for school leader *i*, and $\in_{ci}$ is a random error term. Including the school leader fixed effects in the regression effectively adjusted for differences in the school leaders evaluated on different components. Therefore, differences in the estimates of α_c across different components captured differences in the difficulty of components.

Table E1. Summary statistics on the distribution of Framework for Leadership component scores for principals in the 2012/13 pilot year

	Percentage of principals earning:					
Component	**Failing**	**Needs improvement**	**Proficient**	**Distinguished**	**Average**	**Standard deviation**
1a: Strategic goals	0.0	5.0	73.9	21.1	2.2	0.5
1b: Data for decisionmaking	0.0	5.8	68.1	26.1	2.2	0.5
1c: Empowering work environment	0.0	5.7	66.7	27.6	2.2	0.5
1d: Continuous improvement	0.0	6.5	66.4	27.1	2.2	0.5
1e: Lessons from accomplishments and failures	0.0	2.3	72.2	25.5	2.2	0.5
2a: Leverages resources	0.0	3.4	78.6	18.0	2.1	0.4
2b: School safety	0.3	2.3	63.3	34.0	2.3	0.5
2c: Complies with mandates	0.0	1.5	77.4	21.1	2.2	0.4
2d: Clear expectations for students and staff	0.4	3.6	70.0	26.1	2.2	0.5
2e: Communicates effectively	0.0	7.9	69.8	22.3	2.1	0.5
2f: Manages conflict	0.0	5.7	73.9	20.3	2.1	0.5
3a: School improvement initiatives	0.0	6.1	71.5	22.4	2.2	0.5
3b: Aligns curricula and instruction	0.0	6.0	72.7	21.3	2.2	0.5
3c: High-quality instruction	0.0	11.1	68.5	20.4	2.1	0.6
3d: High expectations for students	0.0	4.0	69.7	26.3	2.2	0.5
3e: Maximizes instructional time	0.0	2.9	68.8	28.3	2.3	0.5
4a: Parent and community involvement	0.0	9.7	67.0	23.3	2.1	0.6
4b: Professionalism	0.0	2.2	58.3	39.5	2.4	0.5
4c: Supports professional growth	0.0	1.7	66.4	31.8	2.3	0.5
All components	**0.0**	**5.0**	**69.6**	**25.5**	**2.2**	**0.5**

Source: Authors' calculations based on Framework for Leadership 2012/13 pilot evaluation scores provided by the Pennsylvania Department of Education.

Adjusted average scores on the components (table E3) confirm the conclusion drawn from the unadjusted averages: components were generally similar in difficulty. Adjusted average scores ranged from 2.1 to 2.4 for principals and from 1.9 to 2.3 for assistant principals.

Detailed tabulations of the distributions of scores on the full Framework for Leadership and its domains

Because in 2012/13 most school leaders earned component scores of proficient (2 points) or distinguished (3 points), the domain scores and full FFL scores were concentrated primarily in the range of 2–3 points. Histograms of full FFL scores (see figures 3 and 4) show evidence that few school leaders scored below 2. Detailed tabulations substantiate the visual evidence from the histograms (tables E4 and E5).

Table E2. Summary statistics on the distribution of Framework for Leadership component scores for assistant principals in the 2012/13 pilot year

	Percentage of assistant principals earning:					
Component	Failing	Needs improvement	Proficient	Distinguished	Average	Standard deviation
1a: Strategic goals	0.0	3.6	91.1	5.4	2.0	0.3
1b: Data for decisionmaking	0.0	8.1	74.2	17.7	2.1	0.5
1c: Empowering work environment	0.0	1.7	74.1	24.1	2.2	0.5
1d: Continuous improvement	0.0	1.8	87.7	10.5	2.1	0.3
1e: Lessons from accomplishments and failures	0.0	3.5	80.7	15.8	2.1	0.4
2a: Leverages resources	0.0	3.8	84.9	11.3	2.1	0.4
2b: School safety	0.0	0.0	78.1	21.9	2.2	0.4
2c: Complies with mandates	0.0	1.8	87.5	10.7	2.1	0.3
2d: Clear expectations for students and staff	0.0	3.4	76.3	20.3	2.2	0.5
2e: Communicates effectively	0.0	8.8	77.2	14.0	2.1	0.5
2f: Manages conflict	0.0	5.4	75.0	19.6	2.1	0.5
3a: School improvement initiatives	0.0	5.6	83.3	11.1	2.1	0.4
3b: Aligns curricula and instruction	0.0	15.1	75.5	9.4	1.9	0.5
3c: High-quality instruction	0.0	5.1	83.1	11.9	2.1	0.4
3d: High expectations for students	0.0	1.7	81.0	17.2	2.2	0.4
3e: Maximizes instructional time	0.0	6.3	81.0	12.7	2.1	0.4
4a: Parent and community involvement	0.0	12.3	75.4	12.3	2.0	0.5
4b: Professionalism	0.0	0.0	64.3	35.7	2.4	0.5
4c: Supports professional growth	0.0	5.0	80.0	15.0	2.1	0.4
All components	**0.0**	**4.9**	**79.4**	**15.7**	**2.1**	**0.4**

Source: Authors' calculations based on Framework for Leadership 2012/13 pilot evaluation scores provided by the Pennsylvania Department of Education.

Table E3. Average Framework for Leadership component scores, adjusted for differences in the mix of school leaders evaluated on different components of the 2012/13 pilot year

	Adjusted average score	
Component	Principals	Assistant principals
1a: Strategic goals	2.1	2.0
1b: Data for decisionmaking	2.2	2.1
1c: Empowering work environment	2.2	2.2
1d: Continuous improvement	2.2	2.1
1e: Lessons from accomplishments and failures	2.2	2.1
2a: Leverages resources	2.1	2.0
2b: School safety	2.3	2.2
2c: Complies with mandates	2.2	2.1
2d: Clear expectations for students and staff	2.2	2.2
2e: Communicates effectively	2.1	2.0
2f: Manages conflict	2.1	2.1
3a: School improvement initiatives	2.1	2.0
3b: Aligns curricula and instruction	2.1	1.9
3c: High-quality instruction	2.1	2.0
3d: High expectations for students	2.2	2.1
3e: Maximizes instructional time	2.2	2.1
4a: Parent and community involvement	2.1	2.0
4b: Professionalism	2.4	2.3
4c: Supports professional growth	2.3	2.1

Source: Authors' calculations based on Framework for Leadership 2012/13 pilot evaluation scores provided by the Pennsylvania Department of Education.

Table E4. Distribution of principals' scores on the full Framework for Leadership and its domains in the 2012/13 pilot year (percent unless otherwise indicated)

Characteristic of distribution	Full FFL	Domain 1	Domain 2	Domain 3	Domain 4
Based on unrounded domain scores					
Average score	2.2	2.2	2.2	2.2	2.2
Standard deviation of scores	0.3	0.4	0.4	0.4	0.4
Distribution of scores					
Below 0.5	0.0	0.0	0.3	0.0	0.0
At least 0.5, below 1.0	0.0	0.0	0.0	0.0	0.0
At least 1.0, below 1.5	1.2	3.0	1.2	4.5	2.7
At least 1.5, below 2.0	15.8	7.7	11.0	10.1	6.3
Exactly 2.0	25.3	47.0	44.0	46.7	46.4
Above 2.0, below 2.5	40.2	21.7	23.8	20.5	17.0
At least 2.5, below 3.0	14.9	11.3	12.8	8.3	14.0
Exactly 3.0	2.7	9.2	6.8	9.8	13.7
Based on domain scores rounded to whole numbers					
Average score	2.2	2.2	2.2	2.1	2.3
Standard deviation of scores	0.4	0.5	0.4	0.5	0.5
Distribution of scores					
Below 0.5	0.0	0.0	0.3	0.0	0.0
At least 0.5, below 1.0	0.0	0.0	0.0	0.0	0.0
At least 1.0, below 1.5	0.9	3.0	1.2	4.5	2.7
At least 1.5, below 2.0	6.5	0.0	0.0	0.0	0.0
Exactly 2.0	56.5	76.5	78.9	77.4	69.6
Above 2.0, below 2.5	13.1	0.0	0.0	0.0	0.0
At least 2.5, below 3.0	12.2	0.0	0.0	0.0	0.0
Exactly 3.0	10.7	20.5	19.6	18.2	27.7

FFL is the Pennsylvania Department of Education Framework for Leadership.

Source: Authors' calculations based on Framework for Leadership 2012/13 pilot evaluation scores provided by the Pennsylvania Department of Education.

Table E5. Distribution of assistant principals' scores on the full Framework for Leadership and its domains in the 2012/13 pilot year (percent unless otherwise indicated)

Characteristic of distribution	Full FFL	Domain 1	Domain 2	Domain 3	Domain 4
Based on unrounded domain scores					
Average score	2.1	2.1	2.1	2.1	2.1
Standard deviation of scores	0.3	0.3	0.3	0.3	0.3
Distribution of scores					
Below 0.5	0.0	0.0	0.0	0.0	0.0
At least 0.5, below 1.0	0.0	0.0	0.0	0.0	0.0
At least 1.0, below 1.5	2.9	2.9	0.0	2.9	1.4
At least 1.5, below 2.0	13.0	7.2	7.2	14.5	10.1
Exactly 2.0	29.0	55.1	53.6	50.7	49.3
Above 2.0, below 2.5	46.4	27.5	24.6	26.1	24.6
At least 2.5, below 3.0	8.7	4.3	10.1	4.3	10.1
Exactly 3.0	0.0	2.9	4.3	1.4	4.3
Based on domain scores rounded to whole numbers					
Average score	2.1	2.0	2.1	2.0	2.1
Standard deviation of scores	0.3	0.3	0.4	0.3	0.4
Distribution of scores					
Below 0.5	0.0	0.0	0.0	0.0	0.0
At least 0.5, below 1.0	0.0	0.0	0.0	0.0	0.0
At least 1.0, below 1.5	1.4	2.9	0.0	2.9	1.4
At least 1.5, below 2.0	1.4	0.0	0.0	0.0	0.0
Exactly 2.0	72.5	89.9	85.5	91.3	84.1
Above 2.0, below 2.5	15.9	0.0	0.0	0.0	0.0
At least 2.5, below 3.0	7.2	0.0	0.0	0.0	0.0
Exactly 3.0	1.4	7.2	14.5	5.8	14.5

FFL is the Pennsylvania Department of Education Framework for Leadership.

Source: Authors' calculations based on Framework for Leadership 2012/13 pilot evaluation scores provided by the Pennsylvania Department of Education.

Appendix F. Technical details of school and leader value-added models

In this study, school leaders' contributions to student achievement growth were estimated using value-added models (VAMs). These contributions were therefore referred to as the school leaders' value-added. The starting point for estimating school leaders' value-added was to estimate their schools' contributions to student achievement growth, known as school value-added. School value-added estimates were then adjusted to distinguish the leaders' contribution from the influences of other school-specific factors. This appendix provides details of the estimation of both school value-added and school leaders' value-added.

Estimating school value-added

Empirical models. The school VAMs estimated schools' contributions to student achievement growth based on Pennsylvania System of School Assessment (PSSA) scores and Keystone Exam scores in the following subjects, grades, and school years:[6]

- PSSA math: grades 4–8 (2007/08 through 2012/13) and 11 (2009/10 through 2011/12).
- PSSA reading: grades 4–8 (2007/08 through 2012/13) and 11 (2009/10 through 2011/12).
- PSSA science: grades 4 and 8 (2007/08 through 2012/13) and 11 (2009/10 through 2011/12).
- PSSA writing: grades 5 and 8 (2007/08 through 2012/13) and 11 (2009/10 through 2011/12).
- Keystone algebra I, English literature, and biology: all spring scores for students in grade 8 or higher (2012/13).

The following regression equation, estimated separately for each subject-grade-year combination, describes the school VAMs for grade 4–8 students using PSSA outcomes:

$$\text{(F1)} \qquad A_{isy} = \beta' P_{i(y-1)} + \gamma' X_{iy} + \delta' S_{isy} + e_{isy}.$$

In the model, A_{isy} is the assessment score for student i attending school s in year y, expressed as a z-score with mean 0 and standard deviation 1 within each subject-grade-year combination. For example, A_{isy} could be the z-score on the grade 5 PSSA math assessment. The vector $\mathbf{P}_{i(y-1)}$ included variables for student i's prior-year PSSA scores. All the VAMs described by equation F1 included prior-year math and reading scores and, when available, prior-year science and writing scores. The prior-year scores came from the previous grade for most students. However, prior scores for grade repeaters came from the same grade as the outcome variable. The vector $\mathbf{P}_{i(y-1)}$ therefore included separate sets of variables for the prior-year scores of grade nonrepeaters and grade repeaters. The vector $\mathbf{X}_{iy}$ was a set of variables for observed student characteristics and for grade repetition. The coefficients in β and γ were the estimated relationships between students' assessment scores and each respective student characteristic, controlling for the other factors in the model. The variable e_{isy} was the error term.

The vector $S_{\boldsymbol{isy}}$ included a school variable for each school in the VAM that was equal to 1 for students attending the school and 0 otherwise. Students attending multiple schools were included in the model on multiple rows of the dataset, once for each school, and each student-school-year observation had exactly one nonzero element in $S_{\boldsymbol{isy}}$. Weights were used to account for a student's exposure to each school that he or she attended during the school year. A student contributed a total weight of 1, which was split evenly across the schools he or she attended during the year (Hock & Isenberg, 2012). This approach gave less weight to students in calculating a school's value-added when students also attended another school in the same year.

The vector δ was a set of coefficients to be estimated, one for each school in the VAM. Each coefficient in δ identified a school's contribution to student learning—the extent to which the actual achievement of students tended to be above or below what was predicted for an average school. The average value-added score for schools across the state was set equal to 0, but this did not mean that student learning was 0 at the school with the average value-added score. Rather, a positive value-added estimate represented above-average school performance and a negative estimate represented below-average performance. The reference point for determining the average school contribution depended on the sample of schools in the model. Since the models included students and schools across the state, the value-added estimates were calculated relative to the contribution of the average school in Pennsylvania in the grade, subject, and school year covered by the VAM.

The school VAM for grade 11 PSSA outcomes and for Keystone Exam outcomes followed equation F1, except that the baseline scores were students' grade 8 PSSA scores because PSSAs were not administered in consecutive grades at the high school level. The baseline scores for grade 8 students taking Keystone Exams were their prior-year PSSA scores.

Two-step estimation process. The VAMs relied on students' own prior achievement scores as indicators of their academic abilities, but standardized tests are imperfect measures of ability. The measurement error introduced by using prior assessment scores as ability measures causes standard regression techniques to produce biased estimates of school effectiveness. The school VAMs accounted for measurement error by incorporating the test/retest reliability of PSSAs into the regression models directly. This approach, called an errors-in-variables regression, eliminated bias due to known measurement error in students' prior-year tests (Buonaccorsi, 2010). Errors-in-variables regression provided a better estimate of β in equation F1 than would be obtained by ordinary regression.

Two regression steps were needed to estimate the VAMs because of a technical limitation of the errors-in-variables regression approach that does not allow for standard errors that are consistent in the presence of both heteroskedasticity and clustering at the student level to be obtained directly. The first step was to estimate equation F1 separately for each grade-subject-year combination (or assessment-year combination for Keystone Exams) with the errors-in-variables regression correction for measurement error in the baseline scores, based on reliability data for the PSSA published by the Pennsylvania Department of Education. This regression output was used to calculate adjusted outcome scores that net out the contribution of all prior test scores:

(F2a) $$\hat{A}_{isy} = A_{isy} - \beta' P_{i(y-1)} \text{ [for students in grades 4–8]}$$

(F2b) $$\hat{A}_{isy} = A_{isy} - \beta' P_{i(8th\ grade)} \text{ [for students in grades 9–12].}$$

The second step was to use the adjusted outcome in place of the actual score and estimate equation F3 by ordinary least squares separately for each grade-subject-year or assessment-year combination:

(F3) $$\hat{A}_{isy} = \gamma' X_{iy} + \delta' S_{isy} + e_{isy}.$$

The standard errors for the estimates from equation F3 were heteroskedasticity-consistent and clustered at the student level.

Controls for students' prior achievement and background characteristics. The school VAMs accounted for several observable factors, including students' prior test scores and background characteristics. The prior test score controls included students' PSSA scores in all available subjects from either the prior year for grade 4–8 students or grade 8 for older students. Students who repeated a grade were included in the VAMs.[7] The school VAMs for students in grades 4–8 include additional PSSA variables for grade repeaters and a separate grade repetition indicator. The school VAMs for grade 11 students and for students taking the Keystone Exams did not include additional PSSA variables for grade repeaters or a grade repetition indicator since the baseline scores for all students in those VAMs came from the same grade (grade 8).

The outcome and baseline assessments used in each VAM for students who did not repeat a grade are shown in table F1. In the science and writing VAMs, it was not possible to include students' same-subject scores from the prior year because these science and writing PSSAs were not given in consecutive grades. While being able to control for same-subject, prior-year scores is preferable because the school effectiveness estimates would be more precise, excluding these variables did not preclude estimating the VAMs.

The study included in the VAMs all students with a baseline test score in the same subject for math and reading VAMs, in math for science VAMs, or in reading for writing VAMs. Students' other baseline scores were imputed if they were missing.[8] The imputations were

Table F1. Assessments used as outcomes and baselines in the school value-added models, 2012/13

Outcome assessment	Outcome grades	Baseline assessments	Baseline grades
PSSA math	4	PSSA math and reading	3
PSSA math	5	PSSA math, reading, and science	4
PSSA math	6	PSSA math, reading, and writing	5
PSSA math	7	PSSA math and reading	6
PSSA math	8	PSSA math and reading	7
Keystone algebra I	8–12	PSSA math, reading, science, and writing	7, 8
PSSA reading	4	PSSA math and reading	3
PSSA reading	5	PSSA math, reading, and science	4
PSSA reading	6	PSSA math, reading, and writing	5
PSSA reading	7	PSSA math and reading	6
PSSA reading	8	PSSA math and reading	7
Keystone English literature	8–12	PSSA math, reading, science, and writing	7, 8
PSSA writing	5	PSSA math, reading, and science	4
PSSA writing	8	PSSA math and reading	7
PSSA science	4	PSSA math and reading	3
PSSA science	8	PSSA math and reading	7
Keystone biology	8–12	PSSA math, reading, science, and writing	7, 8

PSSA is Pennsylvania System of School Assessment.

Note: Baseline scores for grade repeaters were their prior-year scores in the same grade as the outcome variable. Value-added models using Keystone Exams included students in multiple grades because the exams were end-of-course assessments rather than end-of-grade assessments.

Source: Authors' compilation based on data provided by the Pennsylvania Department of Education.

based on the other prior-year scores, outcome scores, and background characteristics of students who had nonmissing scores.

The VAMs also controlled for observable student background characteristics that are thought to be correlated with academic achievement and outside the control of schools (table F2). All of these measures were used in the teacher VAMs estimated by Walsh and Lipscomb (2013). Including observable student background characteristics improved the likelihood that the VAM estimates could measure the direct contributions of schools to student achievement growth versus other factors. Excluded were factors such as measures of family structure or parent educational attainment that were not collected by the Pennsylvania Information Management System. As in the analysis of Walsh and Lipscomb (2013), the gender and race/ethnicity controls were not meant to set different standards for students but rather to recognize that these variables explained statistically significant portions of the variation in student achievement even after accounting for students' prior test scores and the other factors shown in table F2. To the extent that gender and race/ethnicity represented unobserved factors that differed across students and were outside the control of schools, the VAM estimates would systematically penalize or reward certain schools if these controls were omitted.

The sample characteristics of the school VAMs for 2012/13 are shown in table F3. The first column of data shows the error-adjusted standard deviation of school value-added—a

Table F2. Student background control variables used in the school value-added models, 2012/13

Student background control variable	Definition
Free lunch	Free lunch participation (0 or 1)
Reduced-price lunch	Reduced-price lunch participation (0 or 1)
English language learner student	English language learner student in outcome year (0 or 1)
Specific learning disability	Designation of specific learning disability under IDEA (0 or 1)
Speech or language impairment	Designation of speech or language impairment under IDEA (0 or 1)
Emotional disturbance	Designation of emotional disturbance under IDEA (0 or 1)
Intellectual disability	Designation of intellectual disability under IDEA (0 or 1)
Autism	Designation of autism under IDEA (0 or 1)
Physical/sensory impairment	Designation of hearing impairment, visual impairment, deaf-blindness, or orthopedic impairment under IDEA (0 or 1)
Other impairment	Designation of other health impairment, multiple disabilities, developmental delay, or traumatic brain injury under IDEA (0 or 1)
Mobility	Attended multiple schools during school year (0 or 1)
Grade repeater (grade 4–8 models only)	Repetition of the current grade (0 or 1)
Behind grade	More than 1.5 years older than expected for grade (0 or 1)
Age	Student age in years as of September 1
PSSA-Modified (outcome)	Outcome is a PSSA-Modified score (PSSA outcomes only) (0 or 1)
PSSA-Modified (prior-year score)	Prior-year score is a PSSA-Modified score (0 or 1)
Gender	Male (0 or 1)
Race/ethnicity	Indicators for African American, Hispanic, Asian/Pacific Islander, or other race/ethnicity (0 or 1)

IDEA is Individuals with Disabilities Education Act.

PSSA is Pennsylvania System of School Assessment.

Source: Authors' compilation based on data provided by the Pennsylvania Department of Education.

Table F3. Sample characteristics of school value-added models, 2012/13

Outcome	Error adjusted standard deviation of school value added (in student *z* score units)	Number of students	Number of schools
PSSA math, grade 4	0.18	122,479	1,639
PSSA math, grade 5	0.19	121,765	1,537
PSSA math, grade 6	0.18	124,604	1,120
PSSA math, grade 7	0.18	126,995	895
PSSA math, grade 8	0.15	126,227	887
Keystone algebra I	0.34	347,738	1,254
PSSA reading, grade 4	0.15	122,141	1,639
PSSA reading, grade 5	0.14	121,451	1,537
PSSA reading, grade 6	0.12	124,325	1,119
PSSA reading, grade 7	0.12	126,741	894
PSSA reading, grade 8	0.10	125,856	887
Keystone literature	0.17	229,842	768
PSSA writing, grade 5	0.28	119,947	1,537
PSSA writing, grade 8	0.28	124,877	885
PSSA science, grade 4	0.19	122,250	1,638
PSSA science, grade 8	0.16	125,450	885
Keystone biology	0.22	255,123	775

PSSA is Pennsylvania System of School Assessment.

Note: No PSSAs were administered in grade 11 in 2012/13.

Source: Authors' calculations based on student achievement and background data provided by the Pennsylvania Department of Education.

measure of dispersion in the school value-added estimates net of what would be expected based on sampling error alone—expressed in student *z*-score units. For example, a value of 0.18 indicates that, relative to the school at the 50th percentile of the value-added distribution, the school at the 84th percentile was expected to raise student achievement by 0.18 student-level standard deviation, which is equivalent to lifting the median-achieving student in the state to the 57th percentile. The last two columns show the number of students and schools, respectively, included in each VAM. The table does not include VAMs based on grade 11 PSSAs because those assessments were not given in 2012/13.

Obtaining composite school value-added estimates. After estimating school VAMs separately for each subject-grade-year combination, the study constructed composite measures of a school's value-added in each year based on combining its value-added estimates across different grades and subjects from that year. The study used four composite value-added measures for each school in each year of the data:

- An overall composite that combined all of the value-added estimates across subjects for the school.
- A math composite.
- A reading and writing composite.
- A science composite.

The first step to obtain the composites was to standardize the distributions of all individual school value-added estimates to equalize their variances across grades and subjects.[9] The second step was to combine the standardized school value-added estimates by taking a weighted average of those estimates. The weights were proportional to the number of students contributing to a school's estimates, so that value-added estimates for a particular outcome were given more weight at a school if they were based on more students at the school than other value-added estimates were. Standard errors for the composite estimates were calculated based on the precision of the individual value-added estimates and the covariance between pairs of value-added estimates that included the same groups of students. Any schools with fewer than 10 student equivalents were excluded because estimates for these schools were likely to be imprecise.

Estimating school leader value-added

Although models of school leader effectiveness that compare each school leader with other school leaders who have led the same school in different years impose the fewest assumptions, these models were not appropriate for this study because the FFL scores with which the value-added estimates would be compared were available only for school leaders in the 2012/13 school year. On the other hand, school value-added, which captures the contribution of the entire school to student achievement, could be estimated for all school leaders, as described earlier in this appendix.

However, school value-added is an imperfect measure of a school leader's effectiveness because it also reflects other school-level factors affecting student outcomes, including the lingering effects of previous school leaders (Chiang et al., 2012). Therefore, this study developed a new method for estimating school leader value-added by taking school value-added as the starting point and then making adjustments to account for the lingering influences of previous school leaders and other school-level factors.

Adjusting for the effects of previous school leaders and other school-level factors when estimating the value-added of recently hired school leaders. To measure the value-added of recently hired school leaders (those who began their current positions in 2008/09 or later), regression models were estimated to adjust the current value-added of their schools by controlling for measures of "baseline" school value-added, defined as the same schools' value-added in the year before the school leaders started their current positions. Formally, for school leader l, the dependent variable of the regression model was a composite measure of school value-added in the current year y (SVA_{ly}), with separate models for composite measures based on all subjects combined, math, reading and writing, and science. Regardless of the subjects on which SVA_{ly} was based, the regression model controlled for composite measures of baseline school value-added in math ($MSVA_l$), reading and writing ($RSVA_l$), and science ($SSVA_l$). Controlling for baseline school value-added enabled the model to account for the lingering effects of previous school leaders and other persistent school-level factors beyond the current leaders' control.

In addition, because the school VAM for grade 11 PSSA outcomes and for Keystone Exam outcomes used students' grade 8 PSSA scores as baseline scores, the current value-added of high schools could have reflected, in part, growth that students experienced under the current leaders' predecessors if the current leaders began their positions after the students had already completed one or more years of high school. To account for the possibility that the lingering effects of previous school leaders may have been stronger in high schools than in other schools, the regression model also controlled for an indicator of whether the school leader led a school that offered high school grades in year y ($high_{ly}$) and interaction terms between the high school indicator and every measure of baseline school value-added. The final regression model had the following form:

$$\text{(F4)} \quad SVA_{ly} = \alpha_0 + \alpha_m MSVA_l + \alpha_r RSVA_l + \alpha_s SSVA_l + \alpha_h high_{ly} + \alpha_{hm}(high_{ly} * MSVA_l) + \alpha_{hr}(high_{ly} * RSVA_l) + \alpha_{hs}(high_{ly} * SSVA_l) + \textstyle\sum_{y=9}^{12} \alpha_y Year_y + \varepsilon_{ly}.$$

For each school leader the residual from equation F4 was an estimate of his or her contribution to student achievement growth, adjusted for the effects of previous school leaders and other persistent school-level factors. The estimate captured the degree to which school value-added in the current year exceeded or fell short of a prediction based on the same school's value-added under the previous school leader. Estimated coefficients on the baseline school value-added measures from equation F4—shown separately for elementary/middle and high schools—are provided in table F4 for principals and table F5 for assistant principals.[10] This model assumed that baseline school value-added fully captured the effects of the previous school leader and all other school-specific factors beyond the current

Table F4. Relationship between baseline and current school value-added estimates for principals using subject-specific composite value-added measures

		Coefficient on baseline school value-added in math		Coefficient on baseline school value-added in reading/writing		Coefficient on baseline school value-added in science		
Outcome subject	Tenure in current position (years)	Elementary and middle schools	High schools	Elementary and middle schools	High schools	Elementary and middle schools	High schools	Number of school leaders
All combined	1	0.04	0.09	0.26***	0.44***	0.14***	0.19***	2,553
All combined	2	0.04	0.12**	0.29***	0.21***	0.14***	0.20***	1,587
All combined	3	–0.01	0.11	0.26***	0.15*	0.11***	0.01	1,005
All combined	4	–0.06	0.09	0.25***	0.23	0.10**	–0.02	531
All combined	5	–0.01	na	0.33***	na	0.08*	na	260
Math	1	0.33***	0.46***	0.07*	0.20***	–0.01	0.05	2,552
Math	2	0.30***	0.34***	0.08**	0.08	0.02	0.08	1,587
Math	3	0.26***	0.21**	0.08	0.14	–0.00	–0.05	1,005
Math	4	0.16***	0.20	0.11*	0.15	0.01	0.01	531
Math	5	0.18*	na	0.14	na	0.05	na	260
Reading/writing	1	–0.10***	–0.02	0.48***	0.75***	0.06**	0.07	2,552
Reading/writing	2	–0.07	0.00	0.51***	0.43***	0.06**	0.08	1,587
Reading/writing	3	–0.12**	0.02	0.44***	0.33***	0.04	–0.13**	1,005
Reading/writing	4	–0.11*	0.13	0.40***	0.38	0.01	–0.01	531
Reading/writing	5	–0.00	na	0.52***	na	0.01	na	260
Science	1	–0.19***	–0.14*	–0.02	0.11	0.75***	0.67***	2,544
Science	2	–0.19***	0.07	0.10**	–0.06	0.64***	0.63***	1,585
Science	3	–0.25***	0.10	0.11**	–0.16	0.56***	0.33***	1,001
Science	4	–0.37***	–0.21*	0.20***	–0.02	0.54***	–0.05	530
Science	5	–0.38***	na	0.32**	na	0.42***	na	260

* Significant at $p = .10$; ** significant at $p = .05$; *** significant at $p = .01$.

na is not applicable.

Note: Each coefficient represents the predicted change in current school value-added, expressed in school-level standard deviations, associated with a 1 standard deviation increase in baseline school value-added.

Source: Authors' calculations based on student achievement and background data and school leaders' job assignment data provided by the Pennsylvania Department of Education.

Table F5. Relationship between baseline and current school value-added estimates for assistant principals using subject-specific composite value-added measures

		Coefficient on baseline school value-added in math		Coefficient on baseline school value-added in reading/writing		Coefficient on baseline school value-added in science		
Outcome subject	Tenure in current position (years)	Elementary and middle schools	High schools	Elementary and middle schools	High schools	Elementary and middle schools	High schools	Number of school leaders
All combined	1	–0.00	0.13***	0.29***	0.41***	0.19***	0.22***	1,940
All combined	2	–0.01	0.10*	0.25***	0.36***	0.20***	0.26***	1,012
All combined	3	–0.01	0.12	0.35***	0.17*	0.14***	0.13**	528
All combined	4	–0.15**	–0.02	0.33***	0.30	0.18***	–0.05	182
All combined	5	–0.15	na	0.14	na	0.17**	na	67
Math	1	0.29***	0.47***	0.06	0.21***	0.03	0.04	1,940
Math	2	0.31***	0.35***	–0.01	0.17**	0.07*	0.15***	1,012
Math	3	0.36***	0.20**	0.10	0.16	0.04	0.05	528
Math	4	0.13*	0.29	0.19*	0.10	0.07	0.01	182
Math	5	0.09	na	0.04	na	0.10	na	67
Reading/writing	1	–0.13***	0.01	0.55***	0.66***	0.12***	0.13***	1,940
Reading/writing	2	–0.16**	–0.07	0.49***	0.70***	0.17***	0.13**	1,012
Reading/writing	3	–0.19***	0.14	0.57***	0.32***	0.11***	0.10	528
Reading/writing	4	–0.29***	–0.08	0.53***	0.48	0.15**	–0.16*	182
Reading/writing	5	–0.18	na	0.32***	na	0.11	na	67
Science	1	–0.25***	–0.06	–0.01	0.17	0.81***	0.70***	1,933
Science	2	–0.22***	0.07	0.06	0.06	0.67***	0.71***	1,007
Science	3	–0.25***	0.01	0.25***	–0.07	0.51***	0.39***	526
Science	4	–0.38***	–0.33	0.01	0.14	0.55***	0.17	181
Science	5	–0.56**	na	–0.04	na	0.48***	na	66

* Significant at *p* = .10; ** significant at *p* = .05; *** significant at *p* = .01.

na is not applicable.

Note: Each coefficient represents the predicted change in current school value-added, expressed in school-level standard deviations, associated with a 1 standard deviation increase in baseline school value-added.

Source: Authors' calculations based on student achievement and background data and school leaders' job assignment data provided by the Pennsylvania Department of Education.

school leader's control. It also assumed that the current leader's true effectiveness was uncorrelated with baseline school value-added.

The model controlled for subject-specific measures of baseline school value-added instead of one measure based on all subjects to impose fewer restrictions on the functional form. Equation F4 was estimated separately for school leaders (principals and assistant principals) who had led their schools for one, two, three, four, and five years because the relationships between SVA_{ly} and baseline school value-added could have been different for school leaders with different tenure lengths.

The estimation samples included all school leaders in Pennsylvania with valid estimates of current-year school value-added and subject-specific baseline school value-added. To increase the precision of the estimated coefficients, the regressions pooled together all available data years (2008/09 through 2012/13) from which SVA_{ly} could be obtained. Therefore, year indicators ($Year_y$) were also included. Although all available data years were used to estimate equation F4, only value-added estimates from 2012/13 for school leaders in the pilot were subsequently used to assess the concurrent validity of the FFL (see appendix G).

Because the measures of baseline school value-added in equation F4 were estimates, they had measurement error, which would bias the estimated coefficients on those variables toward 0 if not accounted for. To account for measurement error, each baseline school value-added variable was adjusted by an empirical Bayes "shrinkage" procedure before being used in equation F4, such that the regression coefficient on the adjusted variable would no longer be attenuated. Following Morris (1983), the adjusted estimate for each school was approximately equal to a precision-weighted average of the school's initial value-added estimate and the overall mean of all school value-added estimates, with more precise initial estimates receiving greater weight.[11] Therefore, for schools with relatively imprecise initial estimates based on their own students, the empirical Bayes method effectively produced an estimate based more on the average school. For schools with more precise initial estimates based on their own students, the method put less weight on the estimate for the average school and more weight on the estimate obtained from the school's own students. Finally, the empirical Bayes estimates were recentered to have a mean of 0. The procedure effectively reduced the likelihood that very high or low baseline school value-added estimates were the result of chance error, thereby eliminating the bias in equation F4 that would have stemmed from such errors.

Estimating value-added for school leaders who began their current positions before 2008/09. Because student growth data were available only starting in 2007/08, baseline school value-added of longer-serving leaders—those who began their tenures before 2008/09—could not be estimated. Leaders who had led their school for at least six years as of the end of 2012/13 received full attribution of their school's value-added in 2012/13. This assumed that they had sufficient time to shape their school's value-added so that lingering effects of previous leaders were not relevant.

To test the validity of this assumption, the study estimated a variant of equation F4 in which the dependent variable, SVA_{ly}, consisted of current school value-added based on all subjects combined, and the subject-specific baseline school value-added variables were replaced by a single baseline school value-added variable, $CSVA_l$, that was based on all subjects combined and had undergone the shrinkage procedure. Like equation F4, the model controlled for $high_{ly}$, an indicator of whether the school leader led a school that offered high school grades in year y; ($high_{ly}$ * $CSVA_l$), an interaction term between the high school indicator and the school's baseline school value-added; and year fixed effects. Therefore, as in equation F4, the model allowed the relationship between baseline school value-added and current school value-added to be different for elementary/middle school leaders and high school leaders. The resulting regression equation had the following form:

$$\text{(F5)} \qquad SVA_{ly} = \alpha_0 + \alpha_1 CSVA_l + \alpha_h high_{ly} + \alpha_{hc}(high_{ly} * CSVA_l) + \sum_{y=9}^{12} \alpha_y Year_y + \varepsilon_{ly}.$$

To test the assumption that the lingering effects of previous school leaders would be negligible after the current leaders had served for more than five years, equation F5 was estimated separately for school leaders who had led their schools for one, two, three, four, and five years. If the assumption were valid, α_1 and ($\alpha_1 + \alpha_{hc}$) should decrease monotonically with the current leader's length of service and approach zero. However, α_1 and α_{hc} did not follow a generally decreasing pattern with length of service for both principals and assistant principals and certainly did not approach 0 in any case (table F6). Therefore, the available measure of school leader value-added for longer-serving school leaders was less than ideal, and readers should exercise caution when interpreting results for this group of leaders.

The average value-added of school leaders in the pilot was similar to the average for all school leaders statewide

The value-added estimates of all school leaders statewide were standardized to have a mean of 0 and an error-adjusted standard deviation of 1 (separately for recently hired leaders with different tenure lengths and for longer-serving leaders). Therefore, the extent to which the average value-added of pilot participants differed from 0 indicated how dissimilar pilot participants were relative to all leaders statewide in their contributions to achievement growth. For nearly all groups of leaders and all subjects, the average value-added of pilot participants was statistically indistinguishable from the average value-added of all school leaders statewide (table F7). The only exception was for longer-serving principals in science value-added. However, as previously discussed, the available school leader value-added measures for longer-serving leaders were by no means ideal.

Table F6. Relationship between baseline and current school value-added estimates for school leaders using composite value-added measures that combine all subjects

	Coefficient on baseline school value added		
Type of school leader	Elementary and middle schools	High schools	Number of school leaders
Principals who have led their current school for			
1 year	0.43***	0.70***	2,598
2 years	0.46***	0.49***	1,608
3 years	0.35***	0.28***	1,018
4 years	0.29***	0.34**	541
5 years	0.39***	na	264
Assistant principals who have led their current school for			
1 year	0.49***	0.76***	1,986
2 years	0.47***	0.73***	1,033
3 years	0.50***	0.42***	543
4 years	0.39***	0.22	188
5 years	0.19	na	70

* Significant at *p* = .10; ** significant at *p* = .05; *** significant at *p* = .01.

na is not applicable.

Note: Each coefficient represents the predicted change in current school value-added, expressed in school-level standard deviations, associated with a 1 standard deviation increase in baseline school value-added.

Source: Authors' calculations based on student achievement and background data and school leaders' job assignment data provided by the Pennsylvania Department of Education.

Table F7. Mean and standard deviation of the value-added estimates for school leaders participating in the Framework for Leadership 2012/13 pilot year relative to the statewide distribution of school leaders' value-added estimates

Type of school leader	Outcome subject	Mean relative to statewide average (in school leader standard deviations)	Error-adjusted standard deviation (in school leader standard deviations)	Number of school leaders
Recently hired principals	All combined	–0.05	0.91	188
Recently hired principals	Math	–0.01	0.86	188
Recently hired principals	Reading/writing	–0.08	0.95	188
Recently hired principals	Science	0.04	0.96	188
Longer-serving principals	All combined	0.10	0.83	100
Longer-serving principals	Math	–0.02	0.92	100
Longer-serving principals	Reading/writing	0.01	0.92	100
Longer-serving principals	Science	0.35***	0.80	100
Recently hired assistant principals	All combined	0.03	1.00	49
Recently hired assistant principals	Math	–0.03	0.98	49
Recently hired assistant principals	Reading/writing	0.01	0.93	49
Recently hired assistant principals	Science	0.06	1.15	49

*** Significant at *p* = .01.

Note: Recently hired school leaders began their current positions in 2008/09 or later. Longer-serving school leaders began their current positions before 2008/09.

Source: Authors' calculations based on student achievement and background data and school leaders' job assignment data provided by the Pennsylvania Department of Education.

Appendix G. Technical details and supplementary findings on the relationships between Framework for Leadership scores and school leaders' value-added

Earlier, the report discussed the absence of statistically significant relationships between Framework for Leadership (FFL) scores and school leaders' value-added (see, for example, figure 5). This appendix provides details on the method for estimating these relationships and detailed results of the relationships between total, domain, and component FFL scores and school leaders' value-added.

Estimation model

The relationships between full, domain, and component FFL scores and school leaders' value-added were estimated using a regression equation in which the dependent variable was the FFL score (FFL_l) of school leader *l*, with separate regressions for the full FFL score, each domain score, and each component score. The main explanatory variable was the school leader's value-added estimate (VA_l), adjusted using the same empirical Bayes shrinkage as that described in appendix F. The regression model had the following basic form:

(G1)
$$FFL_l = \beta_0 + \beta_1 VA_l + \varepsilon_l,$$

where β_1 measured the average change in the FFL score (measured in points on the FFL) for a unit change in school leader value-added (measured in standard deviations of school leader value-added) and ε_l was a random error term. A standard two-tailed *t*-test for the null hypothesis that β_1 equaled zero assessed the statistical significance of the relationship between the FFL score and school leader value-added. This model was estimated for school leaders only in the 2012/13 pilot and was estimated separately for longer-serving principals, recently hired principals, and recently hired assistant principals.

The basic model in equation G1 was augmented when the estimation sample consisted of recently hired school leaders. As described in appendix F, the value-added of recently hired leaders was estimated separately for—and was therefore not comparable across—leaders with different tenure lengths. Therefore, equation G1 also controlled for four indicator variables identifying recently hired leaders who had served in their current position for two, three, four, and five years.

The sample sizes in the 2012/13 pilot gave rise to only limited precision for estimating the relationship between school leaders' value-added and their FFL scores. Although the estimated relationships presented in this report are expressed as the regression coefficient (β_1) from equation G1, it is advantageous to consider the correlation coefficient when assessing precision so that the study's precision can be compared with that of prior studies that have estimated correlation coefficients. The correlation coefficient between VA_l and FFL_l is just a simple transformation of β_1—specifically, it is equal to β_1 multiplied by the ratio of the standard deviations of the two variables. With the sample sizes in the 2012/13 pilot, the study could have reliably (with 80 percent power) detected a correlation between VA_l and FFL_l if the true correlation was at least 0.20 for recently hired principals, 0.27 for longer-serving principals, 0.38 for recently hired assistant principals, and 0.66 for longer-serving assistant principals. By comparison, prior research found a correlation of 0.24 between the Framework for Teaching and teachers' value-added in Pennsylvania (Walsh

& Lipscomb, 2013). Therefore, for each group of principals, the correlation between FFL scores and value-added would have been reliably detectable only if it had approximately reached the magnitude of the correlation between the Framework for Teaching and teachers' value-added. The correlation would have needed to be even higher for assistant principals—in fact, unrealistically high for longer-serving assistant principals.[12]

Detailed results

Tables G1–G5 contain detailed regression results of various versions of equation G1 where the dependent variable can be total, domain, or component FFL scores and the estimation samples are school leaders who have different tenure lengths and who lead schools of different grade spans. In these tables, β_1 is expressed as the difference in FFL scores between leaders at the 84th and 50th percentile of leader value-added. This is because a unit increase in school leader value-added—an increase of one standard deviation of school leader value-added—is equivalent to moving a leader previously at the 50th percentile to the 84th percentile of the value-added distribution.

Table G1. Association between the Framework for Leadership scores in the 2012/13 pilot year and the value-added estimates for recently hired principals

		Predicted difference in FFL score between principals at 84th and 50th percentiles of value-added	
Outcome	Value-added measure	Estimate	*p*-value
Full Framework for Leadership score	All subjects	0.00	0.917
Full Framework for Leadership score	Math	0.02	0.497
Full Framework for Leadership score	Reading/writing	–0.01	0.811
Full Framework for Leadership score	Science	0.01	0.705
Score on domain 1: Strategic/cultural leadership	All subjects	–0.01	0.875
Score on domain 1: Strategic/cultural leadership	Math	0.01	0.817
Score on domain 1: Strategic/cultural leadership	Reading/writing	0.00	0.897
Score on domain 1: Strategic/cultural leadership	Science	0.00	0.966
Score on domain 2: Systems leadership	All subjects	–0.01	0.756
Score on domain 2: Systems leadership	Math	0.02	0.662
Score on domain 2: Systems leadership	Reading/writing	–0.02	0.568
Score on domain 2: Systems leadership	Science	–0.01	0.814
Score on domain 3: Leadership for learning	All subjects	0.03	0.510
Score on domain 3: Leadership for learning	Math	0.04	0.401
Score on domain 3: Leadership for learning	Reading/writing	0.01	0.761
Score on domain 3: Leadership for learning	Science	0.03	0.415
Score on domain 4: Professional and community leadership	All subjects	0.01	0.841
Score on domain 4: Professional and community leadership	Math	0.03	0.424
Score on domain 4: Professional and community leadership	Reading/writing	–0.01	0.761
Score on domain 4: Professional and community leadership	Science	0.03	0.483

FFL is the Pennsylvania Department of Education Framework for Leadership.

Note: Recently hired school principals began their current position in 2008/09 or later; *n* = 188.

Source: Authors' calculations based on Framework for Leadership 2012/13 pilot evaluation data, student achievement and background data, and school leaders' job assignment data provided by the Pennsylvania Department of Education.

Table G2. Association between the Framework for Leadership scores in the 2012/13 pilot year and the value-added estimates for longer serving principals

Outcome	Value-added measure	Predicted difference in FFL score between principals at 84th and 50th percentiles of value-added	
		Estimate	*p*-value
Full Framework for Leadership score	All subjects	–0.04	0.486
Full Framework for Leadership score	Math	–0.04	0.340
Full Framework for Leadership score	Reading/writing	–0.03	0.538
Full Framework for Leadership score	Science	0.01	0.880
Score on domain 1: Strategic/cultural leadership	All subjects	–0.05	0.383
Score on domain 1: Strategic/cultural leadership	Math	–0.06	0.235
Score on domain 1: Strategic/cultural leadership	Reading/writing	–0.04	0.492
Score on domain 1: Strategic/cultural leadership	Science	0.00	0.983
Score on domain 2: Systems leadership	All subjects	–0.01	0.807
Score on domain 2: Systems leadership	Math	–0.04	0.327
Score on domain 2: Systems leadership	Reading/writing	0.00	0.924
Score on domain 2: Systems leadership	Science	0.04	0.411
Score on domain 3: Leadership for learning	All subjects	–0.06	0.358
Score on domain 3: Leadership for learning	Math	–0.02	0.648
Score on domain 3: Leadership for learning	Reading/writing	–0.06	0.300
Score on domain 3: Leadership for learning	Science	–0.02	0.702
Score on domain 4: Professional and community leadership	All subjects	–0.02	0.661
Score on domain 4: Professional and community leadership	Math	–0.04	0.384
Score on domain 4: Professional and community leadership	Reading/writing	–0.02	0.738
Score on domain 4: Professional and community leadership	Science	0.02	0.771

FFL is the Pennsylvania Department of Education Framework for Leadership.

Note: Longer serving school principals began their current position before 2008/09; *n* = 100.

Source: Authors' calculations based on Framework for Leadership 2012/13 pilot evaluation data, student achievement and background data, and school leaders' job assignment data provided by the Pennsylvania Department of Education.

Table G3. Association between the Framework for Leadership scores in the 2012/13 pilot year and the value-added estimates for recently hired assistant principals

		Predicted difference in FFL score between assistant principals at 84th and 50th percentiles of value-added	
Outcome	Value-added measure	Estimate	*p*-value
Full Framework for Leadership score	All subjects	0.04	0.467
Full Framework for Leadership score	Math	0.03	0.556
Full Framework for Leadership score	Reading/writing	0.03	0.575
Full Framework for Leadership score	Science	0.02	0.634
Score on domain 1: Strategic/cultural leadership	All subjects	0.07	0.306
Score on domain 1: Strategic/cultural leadership	Math	0.03	0.569
Score on domain 1: Strategic/cultural leadership	Reading/writing	0.05	0.387
Score on domain 1: Strategic/cultural leadership	Science	0.06	0.281
Score on domain 2: Systems leadership	All subjects	0.04	0.392
Score on domain 2: Systems leadership	Math	0.05	0.297
Score on domain 2: Systems leadership	Reading/writing	0.02	0.730
Score on domain 2: Systems leadership	Science	0.00	0.921
Score on domain 3: Leadership for learning	All subjects	0.07	0.390
Score on domain 3: Leadership for learning	Math	0.05	0.396
Score on domain 3: Leadership for learning	Reading/writing	0.03	0.724
Score on domain 3: Leadership for learning	Science	0.04	0.453
Score on domain 4: Professional and community leadership	All subjects	–0.01	0.905
Score on domain 4: Professional and community leadership	Math	–0.03	0.554
Score on domain 4: Professional and community leadership	Reading/writing	0.02	0.697
Score on domain 4: Professional and community leadership	Science	–0.01	0.840

FFL is the Pennsylvania Department of Education Framework for Leadership.

Note: Recently hired assistant principals began their current position in 2008/09 or later; $n = 49$.

Source: Authors' calculations based on Framework for Leadership 2012/13 pilot evaluation data, student achievement and background data, and school leaders' job assignment data provided by the Pennsylvania Department of Education.

Table G4. Predicted difference in component scores on the Framework for Leadership in the 2012/13 pilot year between school leaders at 84th and 50th percentiles of value-added estimates

	Recently hired principals		Longer-serving principals		Recently hired assistant principals	
Component	Estimate	*p*-value	Estimate	*p*-value	Estimate	*p*-value
1a: Strategic goals	0.01	0.796	–0.11	0.164	0.07	0.206
1b: Data for decisionmaking	0.08	0.135	–0.03	0.657	–0.03	0.791
1c: Empowering work environment	0.06	0.322	–0.03	0.673	0.10	0.418
1d: Continuous improvement	–0.04	0.358	–0.11	0.251	0.08	0.126
1e: Lessons from accomplishments and failures	0.00	0.951	–0.02	0.824	0.10	0.278
2a: Leverages resources	0.04	0.388	–0.08	0.231	0.03	0.593
2b: School safety	–0.02	0.599	–0.01	0.813	–0.09	0.156
2c: Complies with mandates	0.01	0.844	0.01	0.922	0.04	0.211
2d: Clear expectations for students and staff	–0.05	0.389	–0.03	0.671	0.08	0.376
2e: Communicates effectively	–0.01	0.901	–0.07	0.357	0.09	0.422
2f: Manages conflict	–0.03	0.573	–0.02	0.689	0.16*	0.060
3a: School improvement initiatives	0.10**	0.044	–0.02	0.804	0.05	0.659
3b: Aligns curricula and instruction	0.02	0.678	–0.05	0.510	0.04	0.715
3c: High-quality instruction	0.04	0.458	0.01	0.943	0.10	0.383
3d: High expectations for students	0.08	0.116	–0.07	0.367	0.12	0.180
3e: Maximizes instructional time	–0.02	0.724	–0.04	0.515	0.10	0.108
4a: Parent and community involvement	0.02	0.735	–0.04	0.576	0.02	0.856
4b: Professionalism	–0.01	0.912	0.02	0.768	–0.03	0.744
4c: Supports professional growth	0.04	0.365	–0.08	0.277	0.02	0.681

* Significant at *p* = .10; ** significant at *p* = .05.

Note: Analyses are based on a value-added measure that combines all subjects. Recently hired school leaders began their current position in 2008/09 or later. Longer-serving school leaders began their current position before 2008/09.

Source: Authors’ calculations based on Framework for Leadership 2012/13 pilot evaluation data, student achievement and background data, and school leaders’ job assignment data provided by the Pennsylvania Department of Education.

Table G5. Association between the Framework for Leadership scores in the 2012/13 pilot year and the value-added estimates for principals, by grade span

		Predicted difference in FFL score between principals at 84th and 50th percentiles of value added		
Grade span[a]	Outcome	Estimate	*p*-value	Number of principals
Elementary	Full Framework for Leadership score	–0.01	0.876	120
Elementary	Score on domain 1: Strategic/cultural leadership	–0.02	0.745	120
Elementary	Score on domain 2: Systems leadership	–0.01	0.853	120
Elementary	Score on domain 3: Leadership for learning	0.02	0.739	120
Elementary	Score on domain 4: Professional and community leadership	–0.02	0.701	120
Middle	Full Framework for Leadership score	–0.01	0.885	69
Middle	Score on domain 1: Strategic/cultural leadership	–0.04	0.558	69
Middle	Score on domain 2: Systems leadership	0.05	0.539	69
Middle	Score on domain 3: Leadership for learning	–0.02	0.709	69
Middle	Score on domain 4: Professional and community leadership	–0.02	0.752	69
High	Full Framework for Leadership score	–0.03	0.535	99
High	Score on domain 1: Strategic/cultural leadership	–0.05	0.315	99
High	Score on domain 2: Systems leadership	–0.04	0.491	99
High	Score on domain 3: Leadership for learning	–0.02	0.660	99
High	Score on domain 4: Professional and community leadership	0.00	0.946	99

FFL is the Pennsylvania Department of Education Framework for Leadership.

Note: Analyses are based on a value-added measure that combines all subjects, and the analysis sample consists of all principals participating in the 2012/13 pilot year who have a value-added measure.

a. Elementary schools are defined as those with no grade above 6; middle schools are defined as those with at least one grade above 6 but no grades above 8; high schools are defined as those with at least one grade above 8.

Source: Authors' calculations based on Framework for Leadership 2012/13 pilot evaluation data, student achievement and background data, and school leaders' job assignment data provided by the Pennsylvania Department of Education.

Notes

The authors are grateful to staff at the Pennsylvania Department of Education for their help throughout this study. Robert Holbrook and David Volkman provided much useful background information on the FFL. Jen Whittle supplied data from the FFL pilot evaluations. Milad Elhadri, John Nau, David Ream, and Thresa Stafford provided administrative data on students and school leaders. In its design phase, this study benefited from input from members of the Technical Working Group of the Mid-Atlantic Regional Education Laboratory, including Laura Hamilton, Chris Hulleman, Andy Porter, and Chris Rhoads. Geoffrey Borman and Patrick Wolf provided a helpful review of both the study proposal and an early draft of this report. This study would also not have been possible without the assistance of various individuals at Mathematica Policy Research. Irina Cheban, Serge Lukashanets, and Clare Wolfendale provided excellent statistical programming support. Matt Johnson provided a thoughtful review of early drafts of this report, and Allie Clifford gave production support.

1. Measures of student achievement include value-added assessment system data; student participation in advanced placement courses; student performance on assessments, projects, and portfolios; and student graduation, promotion, and attendance rates.
2. Throughout this report, the FFL's validity refers to the validity of using FFL scores to identify effective and ineffective school leaders.
3. For comparison, all four domains of the Framework for Teaching in Pennsylvania had acceptable internal consistency, with α values ranging from 0.72 to 0.78 (Walsh & Lipscomb, 2013).
4. Results for longer-serving assistant principals are not presented in this report because too few (12) school leaders belonged to this group.
5. Assistant principals were not further divided into grade span subgroups due to the small sample size.
6. The school VAMs based on PSSA scores also included PSSA-Modified (PSSA-M) scores for students with disabilities who were eligible to take modified assessments as a result of their individualized education program.
7. Students with very rare grade progressions—for example, students who appeared to move into a lower grade—were excluded from the VAMs.
8. Missing values of the student characteristics in $\boldsymbol{X}_{iy}$ were also imputed.
9. The process for standardizing the individual VAM estimates involved first mean-centering the estimates and then dividing the mean-centered estimates and their standard errors by the error-adjusted standard deviation of each estimate distribution.
10. For a given baseline value-added measure, the estimated coefficient for high schools was computed as the sum of the coefficient on the baseline value-added measure and the coefficient on the interaction between that measure and the high school indicator.
11. In Morris (1983), because of a correction for bias, the empirical Bayes estimate does not exactly equal the precision-weighted average of the two values. This adjustment increases the weight on the overall mean by $(K - 3)/(K - 1)$, where K is the number of schools. The study incorporates this correction into the shrinkage procedure.
12. By contrast, with the sample sizes projected for the 2013/14 pilot phase (1,170 principals and 507 assistant principals in total), the minimum detectable correlation would be 0.15 or lower for recently hired principals, longer-serving principals, and recently hired assistant principals, assuming the same proportional distribution of leaders into tenure length groups as that observed in 2012/13. The minimum detectable correlation would still be high (0.29) for longer-serving assistant principals.

References

Branch, G., Hanushek, E., & Rivkin, S. (2012). *Estimating the effect of leaders on public sector productivity: The case of school principals.* Working Paper No. 17803. Cambridge, MA: National Bureau of Economic Research. http://eric.ed.gov/?id=ED529199

Buonaccorsi, J. P. (2010). *Measurement error: Models, methods, and applications*. Boca Raton, FL: Chapman & Hall/CRC.

Chiang, H., Lipscomb, S., & Gill, B. (2012). *Is school value-added indicative of principal quality?* Working paper. Cambridge, MA: Mathematica Policy Research.

Coelli, M., & Green, D. (2012). Leadership effects: School principals and student outcomes. *Economics of Education Review*, *31*(1), 92–109. http://eric.ed.gov/?id=EJ953968

Condon, C., & Clifford, M. (2012). *Measuring principal performance: How rigorous are commonly used principal performance assessment instruments?* Washington, DC: American Institutes for Research.

Covay, E., Porter, A., Murphy, J., Goldring, E., Cravens X., & Elliot, S. (2013). *A known group analysis validity study of the Vanderbilt Assessment of Leadership in Education.* Working paper. East Lansing, MI: Michigan State University.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334.

de Vaus, D. A. (2002). *Surveys in social research* (5th ed.). Crows Nest, Australia: Allen & Unwin.

Dhuey, E., & Smith, J. (2012a). *How important are school principals in the production of student achievement?* Working paper. Toronto, ON: University of Toronto.

Dhuey, E., & Smith, J. (2012b). *How school principals influence student learning.* Working paper. Toronto, ON: University of Toronto. http://eric.ed.gov/?id=ED535648

Goldring, E., Cravens, X. C., Murphy, J., Porter, A. C., Elliott, S. N., & Carson, B. (2009). The evaluation of principals: What and how do states and urban districts assess leadership? *The Elementary School Journal, 110*(1), 19–39. http://eric.ed.gov/?id=EJ851761

Goldring, E., Cravens, X., Murphy, J., Porter, A., & Elliot, S. (2012). *The convergent and divergent validity of the Vanderbilt Assessment of Leadership in Education (VAL-ED): Instructional leadership and emotional intelligence.* Working paper. Nashville, TN: Vanderbilt University.

Grissom, J. A., Kalogrides, D., & Loeb, S. (2012). *Using student test scores to measure principal performance.* Working paper. Cambridge, MA: National Bureau of Economic Research.

Hock, H., & Isenberg, E. (2012). *Methods for accounting for co-teaching in value-added models.* Working paper. Washington, DC: Mathematica Policy Research. http://eric.ed.gov/?id=ED533144

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback on teaching: Combining high-quality observations with student surveys and achievement gains.* MET Project Research Paper. Seattle, WA: Bill & Melinda Gates Foundation. http://eric.ed.gov/?id=ED540960

Lipscomb, S., Chiang, H., & Gill, B. (2012). *Value-added estimates for phase 1 of the Pennsylvania Teacher and Principal Evaluation Pilot: Full report.* Cambridge, MA: Mathematica Policy Research. http://eric.ed.gov/?id=ED531795

Milanowski, A., & Kimball, S. (2012). *The relationship between standards-based principal performance evaluation ratings and school value-added: Evidence from two districts.* Rockville, MD: Westat.

Morris, C. N. (1983). Parametric empirical Bayes inference: Theory and applications. *Journal of the American Statistical Association, 78*(381), 47–55.

Porter, A. C., Murphy, J., Goldring, E., Elliott, S. N., Polikoff, M. S., & May, H. (2008). *Vanderbilt Assessment of Leadership in Education: Technical manual.* New York: Wallace Foundation.

Porter, A. C., Polikoff, M. S., Goldring, E., Murphy, J., Elliott, S. N., & May, H. (2010). Investigating the validity and reliability of the Vanderbilt Assessment of Leadership in Education. *The Elementary School Journal*, *111*(2), 282–313. http://eric.ed.gov/?id=EJ913211

Walsh, E., & Lipscomb, S. (2013). *Classroom observations from phase 2 of the Pennsylvania Teacher Effectiveness Pilot: Assessing internal consistency, score variation, and relationships with value added.* Cambridge, MA: Mathematica Policy Research.

The Regional Educational Laboratory Program produces 7 types of reports

Making Connections
Studies of correlational relationships

Making an Impact
Studies of cause and effect

What’s Happening
Descriptions of policies, programs, implementation status, or data trends

What’s Known
Summaries of previous research

Stated Briefly
Summaries of research findings for specific audiences

Applied Research Methods
Research methods for educational settings

Tools
Help for planning, gathering, analyzing, or reporting data or research