Making Connections

January 2016

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

Moira McCullough
Stephen Lipscomb
Hanley Chiang
Brian Gill
Irina Cheban
Mathematica Policy Research

Key findings

- Most school leaders received scores in the top two performance categories in each practice measured by the Framework for Leadership.
- School leaders who received a higher score in one category of leadership practices tended to receive a higher score in the other categories.
- School leaders' scores in one year were moderately consistent (correlation coefficient of .54) with their scores in the next year.
- Principals with larger estimated contributions to student achievement growth (value-added) scored higher overall and on multiple components and domains than principals with lower 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

U.S. Department of Education
John B. King, Jr., *Acting Secretary*

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

National Center for Education Evaluation and Regional Assistance
Joy Lesnick, *Acting Commissioner*
Amy Johnson, *Action Editor*
Felicia Sanders, *Project Officer*

REL 2016–106

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.

January 2016

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.

McCullough, M., Lipscomb, S., Chiang, H., Gill, B., and Cheban, I. (2016). *Measuring school leaders' effectiveness: Final report from a multiyear pilot of Pennsylvania's Framework for Leadership* (REL 2016–106). 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. Those that are doing so face a substantial challenge: there is scant evidence on the validity and reliability of current principal evaluation tools.

Pennsylvania is among states that are developing a new tool for evaluating principals and assistant principals (collectively referred to as school leaders). State legislation passed in 2012 mandates that half a school leader's annual evaluation rating be based on a supervisor's assessment of the quality of leadership practices and that half 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 20 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 school leaders, in preparation for introducing it statewide in 2014/15.

Regional Educational Laboratory (REL) Mid-Atlantic and the Pennsylvania Department of Education (a member of REL Mid-Atlantic's Principal Evaluation Research Alliance) worked together to compile statistical evidence on how well FFL scores measure school leaders' effectiveness. The findings are presented in two reports. An interim report examined the FFL using data from the 2012/13 pilot year (Teh, Chiang, Lipscomb, & Gill, 2014). This final report uses data primarily from the 2013/14 pilot evaluations for 517 principals and 123 assistant principals in Pennsylvania to examine four key FFL properties:

- *Score variation:* the degree to which scores differ across school leaders, which determines whether the FFL can distinguish high- and low-performing school leaders.
- *Internal consistency:* the degree to which different parts of the FFL come to similar conclusions about a school leader's effectiveness. Internal consistency is desirable because the leadership qualities captured by different parts of the FFL are supposed to reflect an overall construct of school leadership ability.
- *Score stability:* the degree to which the same school leader's scores are consistent from one year to the next. Stability helps confirm that FFL is a reliable measure of performance.
- *Concurrent validity:* the degree to which FFL scores in a given year correlate with another measure of school leaders' performance in raising student achievement from the same year.

The following are the key findings of the study:

- Most school leaders received scores of proficient or distinguished (the top two of four performance categories) in each practice measured by the FFL. On average across all components, 95 percent of principals and 96 percent of assistant principals participating in the 2013/14 pilot year scored in the top two performance categories.
- The FFL had good internal consistency for principals (Cronbach's alpha of .90) and acceptable internal consistency for assistant principals (Cronbach's alpha of .79). School leaders who received a higher score in one category of leadership practices tended to receive a higher score in the other categories.

- School leaders' scores in one year were moderately consistent (correlation coefficient of .54) with their scores in the next year. Year-to-year correlations in full FFL scores were similar to those reported for teacher observation instruments by other researchers.
- Principals with larger estimated contributions to student achievement growth (value-added) scored higher overall and on multiple FFL components and domains than principals with lower estimated contributions. The relationships between principal value-added and FFL scores were detected when domain scores were calculated as unrounded averages of component scores and as rounded averages of component scores. When principals were separated into groups by the grade span of their schools, evidence of a relationship between principals' FFL scores and estimated value-added was found for middle school principals only.

These findings indicate that the FFL is a reliable and potentially valid principal evaluation tool. A key strength is its reliability, as measured by both internal consistency and year-to-year stability. The internal consistency of the full FFL is high: school leaders identified as effective or ineffective on one domain tended to be identified similarly on the other domains. The FFL also exhibits moderate year-to-year score stability, comparable to that of widely used teacher observation instruments with demonstrated validity (Kane and Staiger, 2012).

The 2013/14 pilot year provided the first tentative evidence of the FFL's concurrent validity. Scores differentiate to some extent principals who make larger or smaller contributions to student achievement growth. Full FFL scores and scores in two of the four domains are significantly or marginally significantly ($p < .10$) positively associated with both value-added in all subjects combined and value-added in math. This evidence of the concurrent validity of the FFL sets it apart from other principal evaluation tools (Condon & Clifford, 2012; Goldring et al., 2009; Grissom, Kalogrides, & Loeb, 2015; Milanowski & Kimball, 2012).

One area where additional examination by the Pennsylvania Department of Education may be warranted, particularly during wider implementation of the FFL principal evaluation tool, is score distribution. Most school leaders scored in the upper third of the rating scale even though their average effectiveness, as based on their estimated value-added, was statistically indistinguishable from the average for all principals in Pennsylvania. This suggests that supervisors tend to rate school leaders too leniently, even when scores were of low stakes and had no formal consequences, as was the case during the pilot years examined by the study.

Study findings indicated that the Pennsylvania Department of Education may find it useful to gather evidence on the statistical properties of the FFL as the instrument is implemented widely. Monitoring wider implementation will help confirm whether the FFL is a valid and reliable measure of performance across all school leaders in the state—not just among those participating in the pilot. Also, continuing to gather evidence will enable the Pennsylvania Department of Education to examine additional measures of validity and reliability and to refine the FFL as needed.

Contents

Figures

Tables

Why this study?

States and districts across the country are revising how they evaluate school principals. Development and implementation of new systems for evaluating principals have been motivated in part by the option to use the new systems to obtain waivers from particular requirements of the No Child Left Behind Act.

Need for accurate evaluation tools

States and districts that are revising their systems for evaluating principals face a substantial challenge: there is scant evidence on the accuracy of current principal 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 measuring the effectiveness of principals). To inform the selection or development of valid and reliable principal evaluation tools, states and districts need more information on how to accurately measure the quality of principals' leadership practices.

To inform the selection or development of valid and reliable principal evaluation tools, states and districts need more information on how to accurately measure the quality of principals' leadership practices

Pennsylvania is among states that are developing a new tool for evaluating principals and assistant principals (collectively referred to as school leaders). According to Act 82 (2012), half a school leader's annual evaluation rating must be based on a supervisor's assessment of the quality of leadership practices and half must be based on measures of student achievement.[1]

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. It specifies 20 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, assistant superintendent, or, for some assistant principals, the school principal. (One component was added following the 2012/13 pilot year, during which 19 components were rated.) 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 worked with Regional Educational Laboratory Mid-Atlantic to compile statistical evidence on how well

FFL scores measure school leaders' effectiveness. In particular, the study team sought evidence on four key FFL properties:

- *Score variation:* the degree to which scores differ across school leaders, which determines whether the FFL can distinguish high- and low-performing school leaders.
- *Internal consistency:* the degree to which different parts of the FFL come to similar conclusions about a school leader's effectiveness.
- *Score stability:* the degree to which the same school leader's scores are consistent from one year to the next. Stability helps confirm that the FFL is a reliable measure of performance.
- *Concurrent validity:* the degree to which FFL scores in a given year correlate with another measure of school leaders' performance in raising student achievement from the same year.

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

What the study examined

The Pennsylvania Department of Education piloted the evaluation tool with selected groups of school leaders in the 2012/13 and 2013/14 school years before introducing it statewide in the 2014/15 school year (see appendix C for a description of the participants, rating procedures, and completeness of data in the 2013/14 pilot year). The pilot evaluations were used only to provide evidence on FFL properties; they had no formal consequences for rated school leaders. The study examining the pilot evaluation data resulted in two reports. An interim report provided findings based on data from the 2012/13 pilot year (Teh et al., 2014). This final report provides findings based on data primarily from the 2013/14 pilot year, although the examination of score stability incorporates data from both pilot years.

The interim report, using data from the 2012/13 pilot year only, examined three of the FFL properties described in this report: score variation, internal consistency, and the relationship of scores with school leaders' contributions to student achievement growth (concurrent validity). The interim report's key findings were as follows:

- Most school leaders received scores of proficient or distinguished in specific leadership practices.
- The full FFL had good internal consistency for both principals and assistant principals. School leaders who received a higher score in one category of leadership practices tended to receive a higher score in the other categories.
- School leaders with larger estimated contributions to student achievement growth did not, on average, receive higher scores than school leaders with smaller estimated contributions to student achievement growth (Teh et al., 2014).

This final report, using data from 517 principals and 123 assistant principals who participated in the 2013/14 pilot year, seeks to verify and expand on the interim report findings on score variation, internal consistency, and concurrent validity using evidence gathered during implementation of the FFL among a larger sample of school leaders participating in the 2013/14 pilot year. The report also examines score stability using data from both the 2012/13 and 2013/14 pilot years. Score variation, internal consistency, and score stability

were examined using descriptive analyses. The relationship of scores with contributions to student achievement growth was examined using correlational analyses (see box 1 for an overview of the study's data and methods and appendixes C–H for more detail).

Descriptive research questions for this report

To what extent do component, domain, and full FFL scores vary across school leaders? The degree of variation in scores is one indication of how well the FFL distinguishes high- and low-performing school leaders. Similar scores across school leaders would be expected if all school leaders were equally effective. However, prior research has revealed clear differences in principals' effectiveness in raising student achievement (Branch, Hanushek, & Rivkin, 2012; Chiang, Lipscomb, & Gill, forthcoming; Coelli & Green, 2012; Dhuey & Smith, 2012, forthcoming). To distinguish high- and low-performing school leaders, FFL scores should thus also differ meaningfully. Confirming this differentiation is therefore a key aim of the study.

What is the internal consistency of the full FFL and its domains? Internal consistency 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 that should be measured consistently across all parts of the tool, so the same leader's scores on different parts should be consistent.

Internal consistency is the only type of reliability the study can examine. 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 scores from different supervisors (inter-rater reliability) or across different but close points in time (test-retest reliability). However, using two years of pilot data, the study can examine year-to-year stability in a leader's scores, which is a consistency measure similar to a measure of test-retest reliability that uses a longer gap in time between scores.

How stable are full FFL, domain, and component scores across years? Year-to-year stability is important to consider because high instability would suggest low reliability. A small amount of instability is not unwarranted; scores might vary somewhat from year to year based on real changes in a leader's effectiveness. However, large fluctuations in scores from one year to the next would raise concerns that the FFL is not a reliable measure of performance, which would imply that scores in any given year should not be used for high-stakes evaluations.

Correlational research question for this report

To what extent do school leaders' FFL scores correlate with their contributions to student achievement growth? Among other leadership qualities, the FFL aims to measure the leadership qualities needed to improve student achievement. School leaders with larger contributions to achievement should, therefore, receive higher scores. The study assessed the FFL's concurrent validity by comparing school leaders' scores with a measure of their contributions to student achievement growth on statewide assessments in the same year.[2]

Box 1. Data and methods

Data

As in the interim report, 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 2013/14 pilot year for 517 principals and 123 assistant principals. Participating school leaders work primarily in districts receiving U.S. Department of Education Race to the Top funds—which were required to participate in the pilot—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 two components from every domain. Although the FFL as implemented during the 2013/14 pilot year included 20 components, participant scores were collected only for the 19 components that were part of the FFL as implemented in the 2012/13 pilot year. Therefore, the analyses use scores only for those 19 components. On average, in both the 2012/13 and 2013/14 pilot years, school leaders were rated on 16 of the 19 components. Since 2014/15 FFL evaluations have required supervisors to assign a domain score based on the preponderance of evidence within a domain, but supervisors in the 2012/13 and 2013/14 pilot years 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 team 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 team to attribute student achievement growth at those schools to the school leaders. The data included all Pennsylvania principals and assistant principals from 2007/08 to 2013/14.

Data on student achievement scores and background characteristics enabled the study team to estimate school leaders' contributions to achievement growth that controlled for students' prior achievement and backgrounds. The data included all Pennsylvania students in grades 3–12 with achievement data available from 2006/07 to 2013/14 and other background data available from 2007/08 to 2013/14. The student achievement data included scores from end-of-grade assessments (the Pennsylvania System of School Assessment), which are administered in grades 3–8 and 11, and end-of-course assessments (the Keystone Exams), which are administered primarily in grades 9–12.

Methods

The methods used in this report to examine score variation, internal consistency, and concurrent validity were consistent with those described in the interim report.

Analyses to address the research question on score variation described the distributions of component, domain, and full FFL scores. The distribution of component scores was characterized by the percentage of school leaders who received 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 D for technical details). The distributions of domain and full

(continued)

Box 1. Data and methods *(continued)*

FFL scores were characterized by the percentage of school leaders in different intervals of the 0–3 point scale.

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

Analyses to address the research question on score stability used data on FFL scores for participants in both pilot years to calculate Pearson's correlation coefficient, a measure of the strength of linear association between scores in each year that ranges from –1 to 1 (see appendix F for technical details). The study team calculated correlations across the 2012/13 and 2013/14 pilot years for full FFL, domain, and component scores of the 189 principals participating in both pilot years.

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 2013/14—referred to as the leaders' value-added. The study team estimated value-added only for recently hired principals—that is, those who began their current leadership roles in 2008/09 or later. For these leaders value-added was estimated as the school's contribution to student achievement growth in 2013/14, adjusted for the same school's contribution under the current leader's predecessor. (The study did not estimate value-added for assistant principals or for principals who began their current roles prior to 2008/09. For the latter group of school leaders, achievement growth data for their predecessors were not available, and thus the necessary adjustments for predecessor contributions could not be made. See appendix G for technical details on estimating value-added.) The final step was to estimate a regression model for the relationship between recently hired principals' FFL scores from the end of the 2013/14 school year and their estimated value-added in the same year (see appendix H for technical details on this model).

What the study found

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

Variation in Framework for Leadership scores was limited, with most component ratings 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 at raising student achievement (Branch et al., 2012; Chiang et al., forthcoming; Coelli & Green, 2012; Dhuey & Smith, 2012, forthcoming). Thus, FFL scores, which are intended to measure performance on leadership components associated with improved student learning, among other outcomes, should vary considerably as well.

Variation in scores was examined at three levels: component, domain, and full FFL. Because only component scores were collected during the pilot, two approaches were used to calculate domain scores. The first approach, used throughout this report, calculates

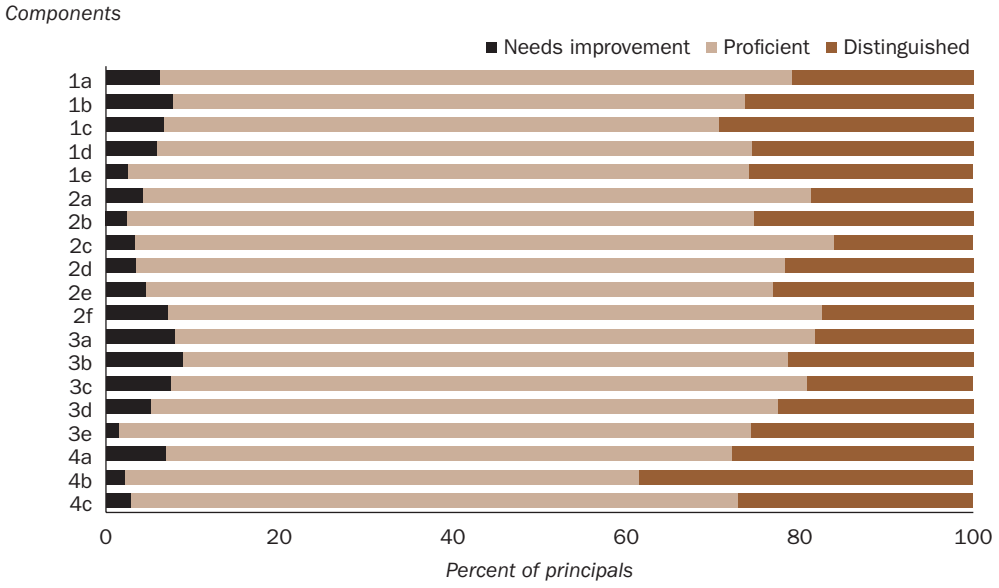
each domain score as the unrounded, equal-weighted average of component scores for the domain (see box 1). Prior to the FFL being widely implemented, the Pennsylvania Department of Education intended for supervisors to assign a domain score that is a whole number. To more closely replicate these anticipated domain scores, the second approach calculates a domain score by rounding the equal-weighted average of component scores within the domain to the nearest whole number. Under both approaches the full FFL score is the equal-weighted average of domain scores, which is how the full FFL score is calculated now that the FFL has been widely implemented in school leader evaluations.

On every component, most principals and assistant principals received a rating of proficient or distinguished. On average, across all components, 95 percent of principals and 96 percent of assistant principals participating in the 2013/14 pilot year were rated either proficient or distinguished—the top two performance categories (figures 1 and 2; see also tables D1 and D2 in appendix D). The proportions of proficient and distinguished component ratings were nearly identical in the 2012/13 pilot year—that is, 95 percent of principals and 95 percent of assistant principals (Teh et al., 2014). The most common rating of performance on any FFL component was proficient (ranging from 59 percent to 80 percent of principals and from 57 percent to 87 percent of assistant principals across components). On average, supervisors of principals assigned the needs improvement rating about 5 percent of the time, and supervisors of assistant principals about 4 percent of the time. Consistent with the 2012/13 pilot year, failing ratings were extremely rare; only four principals and one assistant principal received a failing rating on a component.

On average, across all components, 95 percent of principals and 96 percent of assistant principals participating in the 2013/14 pilot year were rated proficient or distinguished—the top two performance categories

Because school leaders decided jointly with their supervisors which FFL components to use in the pilot evaluations, the set of components on which each school leader was rated

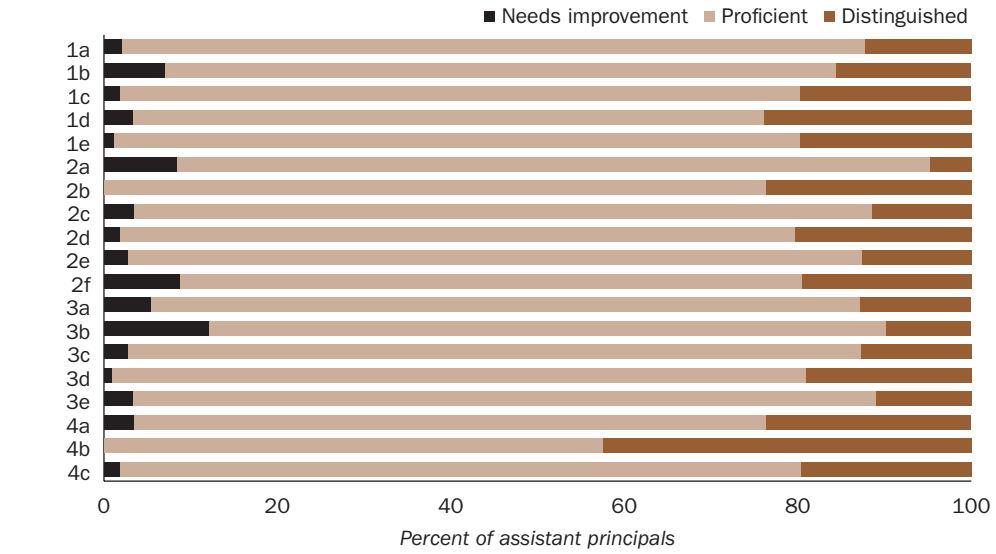
Figure 1. On every component of the Framework for Leadership, principals were most frequently rated as proficient or distinguished in the 2013/14 pilot year



Note: The number of principals receiving a rating was between 398 and 506, depending on the component. Three principals received a failing rating on component 3b, and one principal received a failing rating on component 4b. See table B1 in appendix B for definitions of the components.

Source: Authors' calculations based on Framework for Leadership 2013/14 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 2013/14 pilot year



Note: The number of assistant principals receiving a rating was between 84 and 118, depending on the component. One assistant principal received a failing rating on components 2d and 3c; see table B1 in appendix B for definitions of the components.

Source: Authors' calculations based on Framework for Leadership 2013/14 pilot evaluation scores provided by the Pennsylvania Department of Education.

The findings suggest that allowing school leaders and their supervisors to choose which components to include in the evaluation does not compromise the fairness of the scores

varied. When not every school leader is rated on the same set of components, the relative difficulty of each component may have implications for the fairness of scores across leaders. For the FFL to provide a fair evaluation when supervisors and school leaders choose the components to be rated, as they did for the 2012/13 and 2013/14 pilot evaluations, the difficulty of scoring well should be about the same for each component. As in the 2012/13 pilot year, component score distributions for the 2013/14 pilot evaluations did not differ substantially across components (see figures 1 and 2). For both groups of school leaders, components did not differ systematically in their difficulty after differences in the mix of school leaders evaluated on each component were controlled for (see table D3 in appendix D). The findings suggest that allowing school leaders and their supervisors to choose which components to include in the evaluation does not compromise the fairness of FFL scores.

Scores for each domain and the full FFL were concentrated at the top third of the scale in the 2013/14 pilot year. Consistent with the prevalence of high component score ratings, domain scores assigned to school leaders were overwhelmingly likely to equal 2.0 or above on the 0–3 point scale. In every domain the percentages of both principals and assistant principals scoring at least 2.0 exceeded 85 percent based on unrounded domain scores and 95 percent based on domain scores rounded to whole numbers (see tables D4 and D5 in appendix D).

Likewise, full FFL scores for both principals and assistant principals in the 2013/14 pilot year were concentrated at the top third of the rating scale (figures 3 and 4), which is consistent with the distribution in the 2012/13 pilot year distribution (Teh et al., 2014). With full

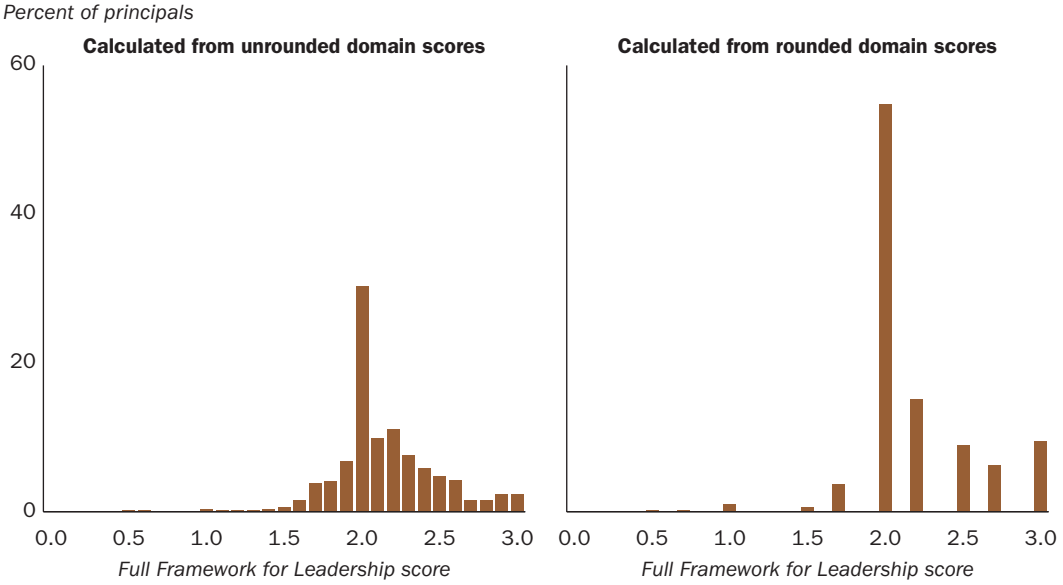
scores calculated from unrounded domain scores, 81 percent of principals and 87 percent of assistant principals had a full score of 2.0 or higher. With full scores calculated from rounded domain scores, 94 percent of principals and 97 percent of assistant principals did. The most common full score was exactly 2.0: 19 percent of principals and 24 percent of assistant principals received this score based on unrounded domain scores, and 55 percent of principals and 58 percent of assistant principals received this score based on rounded domain scores.

Using whole numbers for domain scores reduces score variation. Rounding domain scores to whole numbers—as would be done after supervisors assigned domain scores by judging the preponderance of evidence—lowers the variation in full FFL scores compared with calculating domain scores as unrounded averages of component scores. In both the 2012/13 and 2013/14 pilot years there were fewer distinct values for the full scores when they were calculated from rounded rather than unrounded domain scores (see figures 3 and 4 for 2013/14 distributions and the 2012/13 distributions in Teh et al., 2014). Moreover, because most unrounded domain scores were within 0.5 point of 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 receive a 2 on every domain and thus have the identical full 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.

Compared with calculating domain scores as unrounded averages of component scores, rounding domain scores to whole numbers—as would be done when judging the preponderance of evidence—lowers the variation in full FFL scores so that the FFL could not make any distinctions in performance among a majority of school leaders

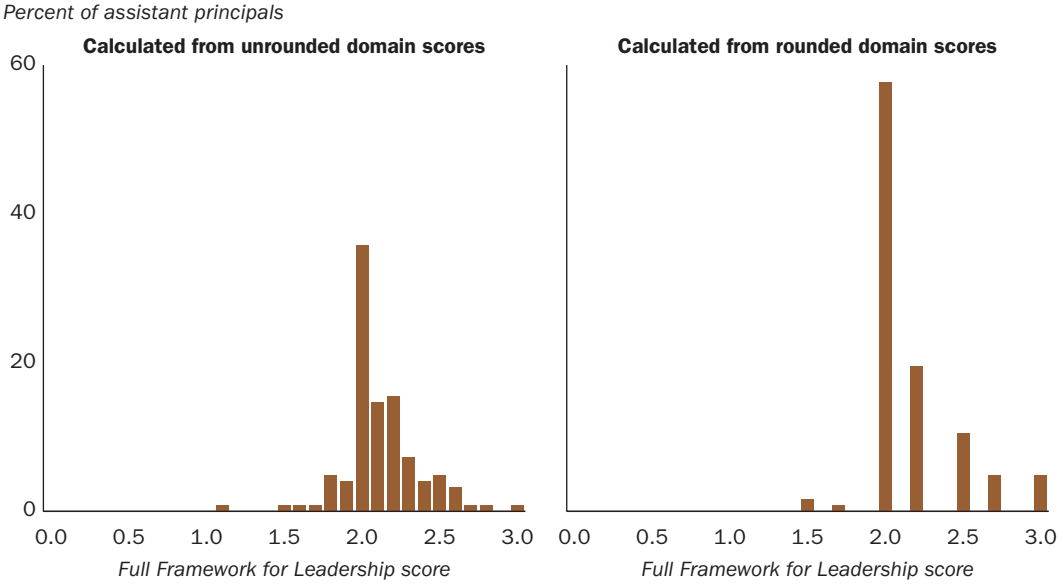
The prevalence of high scores among school leaders could have occurred if highly effective leaders were most likely to participate in the pilot. However, as shown later in this report, leaders participating in the pilot made contributions to student achievement growth that varied substantially and were indistinguishable from the contributions of nonparticipating

Figure 3. Full Framework for Leadership scores were concentrated at the top third of the scale among principals in the 2013/14 pilot year



Source: Authors' calculations based on Framework for Leadership 2013/14 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 2013/14 pilot year



Source: Authors' calculations based on Framework for Leadership 2013/14 pilot evaluation scores provided by the Pennsylvania Department of Education.

The full FFL had good internal consistency for principals and acceptable internal consistency for assistant principals in the 2013/14 pilot year

school leaders, on average. Because there is no evidence that the leaders in the pilot were unusually effective, it appears that supervisors were lenient in assigning ratings.

The full Framework for Leadership had good internal consistency for 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 all the components 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 that 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 (see appendix E). The following critical α values are used in this study:

- .8 or higher is considered good.
- .7 or higher but less than .8 is considered acceptable.
- .6 or higher but less than .7 is considered marginally acceptable.
- Below .6 is considered not acceptable.

The critical values for good and acceptable internal consistency come from a textbook on surveys in social research by de Vaus (2002); a recent analysis of the internal consistency of the Framework for Teaching in Pennsylvania adopted these values as well (Walsh & Lipscomb, 2013). This study follows Teh et al. (2014) and adopts an additional critical value to indicate marginally acceptable internal consistency, because no research suggests that

a .7 value of Cronbach's alpha is a strict threshold for determining whether an evaluation tool should be implemented.

The full FFL had good internal consistency for principals and acceptable internal consistency for assistant principals in the 2013/14 pilot year. The value of Cronbach's alpha was .90 for principals and .79 for assistant principals (table 1). The internal consistency for both types of school leaders had been good in the 2012/13 pilot year, as it was for Pennsylvania teachers evaluated using the Framework for Teaching in the 2011/12 pilot year (Teh et al., 2014; Walsh & Lipscomb, 2013). The full FFL's internal consistency for assistant principals was just below the critical value for good internal consistency in the 2013/14 pilot year and just above this value in the 2012/13 pilot year, a difference that is unlikely to be meaningful. For both types of school leaders, the main conclusion about internal consistency from the 2013/14 pilot year is that the different domains continued to yield similar assessments of a school leader's effectiveness.

The different domains of the FFL continued to yield similar assessments of a school leader's effectiveness

The internal consistency of FFL 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 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 acceptable for domain 1, marginally acceptable for domains 2 and 3, and not acceptable for domain 4.[3] These results are generally consistent with the results found in the first pilot year (Teh et al., 2014).

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 both 2012/13 and 2013/14 pilot years—will distort FFL scores. With an internally consistent measure, 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).

This study cannot determine why internal consistency of the domains was higher for principals than for assistant principals. As with the interim report, this report offers two possible explanations for this pattern. First, superintendents and assistant superintendents, who supplied most of the ratings for both principals and assistant principals (see figure C1 in appendix C), may have had less direct knowledge about assistant principals' performance.

Table 1. The internal consistency of the full Framework for Leadership was good for principals and acceptable for assistant principals in the 2013/14 pilot year

School leader type	Internal consistency[a] (Cronbach s alpha)	Sample size
Principals	.90	517
Assistant principals	.79	123

a. A Cronbach's alpha of .8 or higher is considered good; .7 or higher but less than .8 is considered acceptable; .6 or higher but less than .7 is considered marginally acceptable; below .6 is considered not acceptable.

Source: Authors' calculations based on Framework for Leadership 2013/14 pilot evaluation scores provided by the Pennsylvania Department of Education.

Table 2. The internal consistency of Framework for Leadership domains was higher for principals than for assistant principals in the 2013/14 pilot year

School leader type and Framework for Leadership domain	Internal consistency[a] (Cronbach's alpha)	Sample size
Principals		
Domain 1: Strategic/cultural leadership	.77	386
Domain 2: Systems leadership	.75	369
Domain 3: Leadership for learning	.80	366
Domain 4: Professional and community leadership	.61	388
Assistant principals		
Domain 1: Strategic/cultural leadership	.73	85
Domain 2: Systems leadership	.66	80
Domain 3: Leadership for learning	.67	82
Domain 4: Professional and community leadership	.58	85

a. A Cronbach's alpha of .8 or higher is considered good; .7 or higher but less than .8 is considered acceptable; .6 or higher but less than .7 is considered marginally acceptable; below .6 is considered not acceptable. For each domain, observations are only included in the calculation of internal consistency if a school leader is rated on every component in the domain.

Source: Authors' calculations based on Framework for Leadership 2013/14 pilot evaluation scores provided by the Pennsylvania Department of Education.

The internal consistency of FFL domains provides some assurance against the concern that allowing supervisors and school leaders to choose which components to use in evaluations will distort FFL scores

If so, component scores for assistant principals would be subject to more error and, consequently, would be less consistent. Second, 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. To ensure that FFL scores reflect a coherent assessment of assistant principals' performance, supervisors may need to obtain additional input from colleagues with direct knowledge of that performance and review the position's actual responsibilities before determining which components should factor into the domain scores.

Internal consistency was lowest in domain 4 (professional and community leadership) for both types of school leaders. Findings from the interim report indicated that domain 4, which measures professional and community leadership, may need further development because the components in the domain exhibited the weakest relationship to each other (Teh et al., 2014). In the 2013/14 pilot year Cronbach's alpha for assistant principals in domain 4 improved but was still in the not acceptable range, while the value for principals remained marginally acceptable. Domain 4, therefore, may need further development.

As suggested in the interim report, the three components in domain 4 might not be sufficient; by comparison each of the other domains had five or six components. Adding components to a scale measure typically increases internal consistency by incorporating more information on the underlying concept of interest. As noted in the interim report, professional and community leadership may be distinct concepts, with components in the domain pertaining to one concept or the other but not to the single, collective concept intended to be captured by domain 4 (Teh et al., 2014). The internal consistency of a scale made up of just the two components that pertain to professional leadership (4b and 4c) is marginally acceptable for both types of school leaders, suggesting that the component on community leadership (4a) measures a different school leadership concept (see table E2 in appendix E).

Framework for Leadership scores were moderately stable across years

Score stability indicates the degree to which each school leader's FFL scores are consistent from one year to the next. Wide fluctuations in a school leader's scores from one year to the next could imply that FFL scores are not reliable indicators of effectiveness. At the same time, some instability is acceptable and even anticipated—for example, scores would be expected to increase as school leaders improve over time.

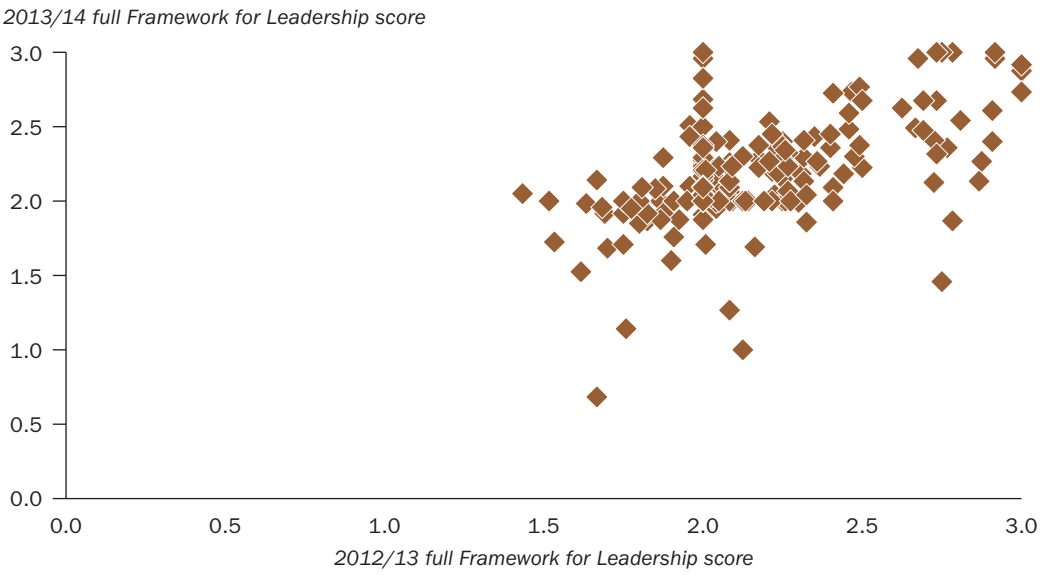
To measure score stability from the 2012/13 pilot year to the 2013/14 pilot year, the study calculates the correlations of the full, domain, and component FFL scores across years among the sample of principals who received ratings in both years. Correlations are calculated as Pearson's correlation coefficients, a statistic that ranges from –1 to 1, where higher positive values are associated with higher stability (see appendix F for more details). The study uses the following standards to categorize the stability of FFL scores using correlation magnitudes (Cohen and Cohen, 1983):

- Correlation coefficient of .8 or higher is considered high stability.
- Correlation coefficient of .6 or higher but less than .8 is considered moderate to high stability.
- Correlation coefficient of .4 or higher but less than .6 is considered moderate stability.
- Correlation coefficient of .2 or higher but less than .4 is considered low stability.

Approximately 54 percent of the variation in full FFL scores across principals represents differences in their effectiveness that persist across years; the remainder of the variation is evident only in individual years and not persistent

Full FFL scores for principals were moderately stable during the two pilot years. (Stability for assistant principals is reported in appendix F because the correlation coefficient is less reliable due to the small sample of assistant principals with two years of scores.) The correlation coefficient was .54 across the full sample of principals participating in both pilot years (figure 5). In other words, 54 percent of the variation in full FFL scores across principals represents differences in their effectiveness that persist across years; the

Figure 5. Full Framework for Leadership scores for principals are moderately stable across years



Source: Authors' calculations, based on Framework for Leadership 2012/13 and 2013/14 pilot evaluation scores provided by the Pennsylvania Department of Education.

remainder of the variation is evident only in individual years and not persistent. A correlation coefficient of .54 is consistent with other findings on the stability of teacher observation instruments (Brophy, Coulter, Crawford, Evertson, & King, 1975; Polikoff, 2015).

Stability of the full FFL score was highest among principals who were rated on the same set of components by the same supervisor in both years. When the sample is limited to principals rated by the same supervisor in both years, the correlation coefficient for full FFL scores is .60. Some of this consistency may be due to a supervisor's preconceived perception of a school leader's effectiveness carrying over from one year to the next and some may be due to consistency in the use of the FFL to assess effectiveness in each year independently. Among the 149 principals rated by the same supervisor, two-thirds were also rated on the same set of components in both years. The consistency of the full score was highest among this group, with a correlation coefficient of .68. In contrast, year-to-year correlation of full scores for principals who had the same rater in both years but were rated on a different set of components in each year is only .38.

The leadership for learning domain had the highest year-to-year stability, and the professional and community leadership domain had the lowest

Among the sample of principals rated by a different supervisor in the 2012/13 and 2013/14 pilot years, the correlation of full FFL scores across years is .40. The correlation is slightly higher (.42) when the group is further limited to principals rated on the same set of components. This correlation coefficient is similar to the observed year-to-year stability of teacher observations for the Measures of Effective Teaching study, which used multiple raters. For example, scores on the Danielson Framework for Teaching had a year-to-year correlation of .44 in that study (Polikoff, 2015).

Year-to-year stabilities for each of the four domains fall in the moderate range. The leadership for learning domain had the highest year-to-year stability (.49), and the professional and community leadership domain had the lowest (.41; see table F1 in appendix F). All correlations for full scores and domain scores were statistically significant at the 5 percent level and have similar magnitudes whether the unrounded or rounded domain scores are used.

Higher Framework for Leadership scores were associated with larger estimated contributions to student achievement growth

One way to assess whether the FFL is working as intended is to examine the relationship between FFL scores and school leaders' contributions to student achievement growth. The Pennsylvania Department of Education regards the leadership practices measured by the FFL as school leaders' key inputs into improving student achievement. If the FFL is indeed measuring these key inputs, FFL scores should be positively related to contributions to student achievement. The study examined correlations of school leaders' 2013/14 scores with an objective measure of their contributions to student achievement growth in the same year. Because both the scores and the objective measure to which they are compared are supposed to capture school leaders' effectiveness in the same school year (2013/14), this analysis provides an assessment of the FFL's concurrent validity. (Cross-year associations between scores and contributions to student achievement growth were also examined as a sensitivity check but used a very limited sample of principals; see appendix H for details.)

This study measures principals' contributions to student achievement growth using a value-added statistical model. The starting point for measuring principals' contributions

is the effectiveness of their schools—captured by how much student achievement growth that year exceeded or fell below predictions based on students' prior achievement and other characteristics. However, a school's value-added may be affected by many factors other than the principal, including the previous principal, the effectiveness of the teachers inherited by the current principal, and community characteristics. To account for these other factors, value-added estimates for principals are measured based on how the school's value-added in 2013/14 deviates from its predicted value-added, which is based on its value-added before the current principal arrived. In other words, the principal's value-added measures how much better or worse the school is performing than it would perform under an average principal, given the school's own prior performance.

Data were not available to measure schools' value-added prior to 2008/09. As a result, the study team could not generate value-added estimates for the 125 longer serving principals. The relationships between value-added and scores were therefore estimated only for principals who began leading their current schools in 2008/09 or later. Assistant principals were not included in the analysis because it is unclear how to isolate their unique contributions to student achievement growth.

Despite the fact that the FFL scores of pilot principals were concentrated at the high end of the scale, the estimated average value-added of pilot participants was statistically indistinguishable from the average of all school leaders in the state

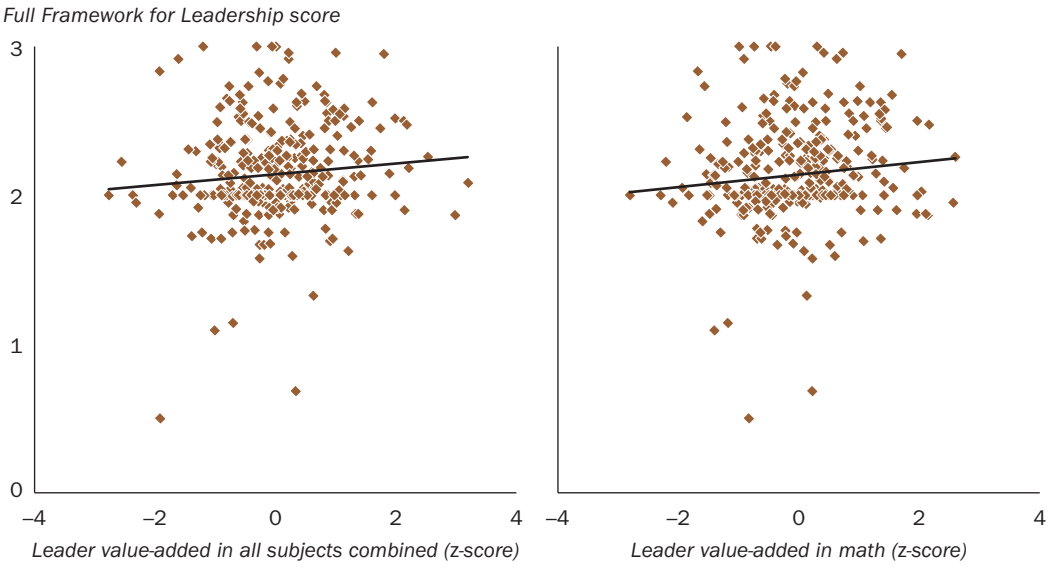
The FFL's concurrent validity could vary depending on whether components, domains, or the full FFL is considered. The domain and component scores with the largest positive associations with value-added could represent promising practices for the Pennsylvania Department of Education to target for professional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

Value-added scores for principals were comparable to those of principals who did not participate in the pilot. Despite the fact that the FFL scores of pilot principals were concentrated at the high end of the scale,[4] the estimated average value-added of pilot participants was statistically indistinguishable from the average of all school leaders in the state (see table G6 and accompanying text in appendix G). These results are similar to those presented in the interim report on the FFL that analyzed the 2012/13 pilot data.

Higher full FFL scores were significantly associated with higher value-added in math and marginally significantly associated with higher value-added in science and in all subjects combined but were not significantly associated with higher value-added in reading or writing. Despite the limited range of FFL scores, principals' full scores had a marginally significant ($p < .10$) positive relationship with estimated value-added in all subjects combined (figure 6; see also table H1 in appendix H). Full scores were also significantly related ($p < .05$) to value-added in math and marginally significantly related to value-added in science (see left panel of figure 6 and table H1 in appendix H). The gently sloping upward lines in the left and right panels of figure 6 indicate that principals with higher estimated contributions to student achievement growth in all subjects combined and in math tended to have higher full FFL scores than principals with lower estimated contributions to student achievement growth. The study found no evidence of a relationship between full FFL scores and estimated value-added in reading or writing.

The magnitudes of the relationships between full FFL scores and value-added measures are small. A principal at the 84th percentile of value-added across all subjects combined is

Figure 6. Higher full Framework for Leadership scores are associated with higher value-added in all subjects combined and in math among recently hired principals



Note: Correlation with all subjects combined is significant at $p < .10$; correlation with math is significant at $p < .05$. Recently hired principals began at their current schools in 2008/09 or later.

Source: Authors' calculations based on Framework for Leadership 2013/14 pilot evaluation data, student achievement and background data, and school leaders' job assignment data provided by the Pennsylvania Department of Education.

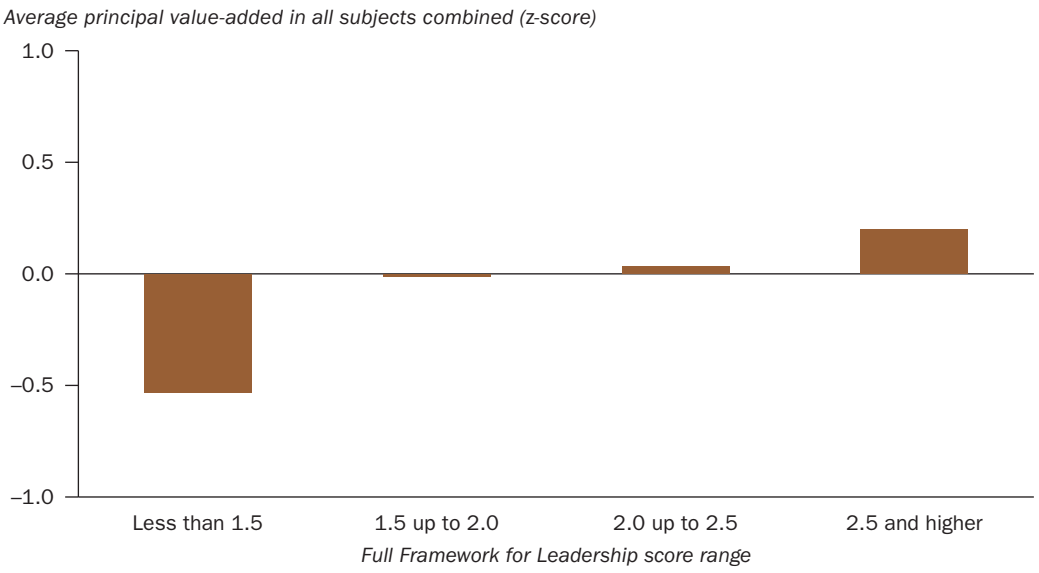
Shows professionalism (component 4b) has the largest relationship of any individual component with estimated value-added across all subjects combined and math

predicted to receive a full score that is only 0.04 higher than the full score of a principal at the 50th percentile. Scores most clearly differentiate among principals in terms of their value-added at the highest and lowest ranges of the scale (figure 7).

Higher FFL domain 2 (systems leadership) and domain 4 (professional and community leadership) scores were associated with larger estimated value-added in all subjects combined. Scores in domains 2 and 4 were also positively related to estimated value-added in math and in science (the relationship between scores in domain 4 and value-added in science was marginally significant). No association with value-added in reading or writing was detected for any domain score. Component 4b (shows professionalism) has the largest relationship of any individual component with estimated value-added across all subjects combined and math (see tables H2 and H3 in appendix H) and is likely driving the relationship between domain 4 scores and estimated value-added.

While no relationships with estimated value-added were detected for FFL domain 1 (strategic and cultural leadership) and domain 3 (leadership for learning) scores, several components in both domains were consistently associated with value-added scores across subjects (see tables H2–H5 in appendix H). Higher scores on components 1b (uses data for informed decisionmaking) and 1c (builds a collaborative and empowering work environment) were related to higher estimated value-added in all subjects combined (at a marginal level of significance) and in math. The magnitudes of the relationships were also among the largest of any components. Scores on components measuring the implementation of high-quality instruction (3c) and the maximizing of instructional time (3e) were also related at a marginal level of significance to value-added in all subjects combined and in math.

Figure 7. Higher full Framework for Leadership score ranges are associated with higher value-added in all subjects combined among recently hired principals



Note: Average principal value-added corresponds to a *z*-score in the principal performance distribution. Recently hired principals began at their current schools in 2008/09 or later.

Source: Authors' calculations based on Framework for Leadership 2013/14 pilot evaluation data, student achievement and background data, and school leaders' job assignment data provided by the Pennsylvania Department of Education.

Among the subset of middle school principals, scores in domain 1 (strategic and cultural leadership) were significantly positively associated with value-added in all subjects combined

Overall, the study found more statistically significant associations between individual components and estimated value-added than would be expected by chance. Nine statistically significant associations were found, while only about four would be expected to occur by chance. Two of 19 components were significantly positively related to value-added across all subjects, five of 19 components were significantly positively related to value-added in math, one of 19 components was significantly positively related to value-added in reading and writing, and one of 19 components was significantly positively related to value-added in science. This finding indicates that a majority of the relationships detected were not spurious and likely reflect true correlations between principals' measured performance on these practices and their estimated contributions to student achievement growth.

Higher FFL scores were associated with larger value-added among middle school principals, but no relationships were detected for elementary school principals or high school principals. Among the subset of middle school principals, scores in domain 1 (strategic and cultural leadership) were significantly positively associated with value-added in all subjects combined (see table H6 in appendix H). Middle school principals' value-added in all subjects combined also had a marginally significant and positive relationship with their full FFL scores and domain 4 scores. The magnitude of the three relationships exceeded the size of all relationships detected across the full sample. No associations between full FFL or domain scores and principal value-added in all subjects combined were detected for either subset of elementary school principals or high school principals. This finding may reflect that value-added estimates typically cover a larger proportion of grades for middle schools than for elementary and high schools and thus are more accurate measures of schoolwide performance. The smaller sample sizes for this analysis, conducted separately by grade span, also made it more difficult to detect statistically significant relationships.

Similar significant relationships between principals' FFL scores and their value-added were detected when rounded domain scores were used. As described earlier, domain scores calculated as rounded averages of component scores most closely replicate the whole number domain scores supervisors might assign in practice. Using these rounded domain scores and associated full FFL score reinforces findings about the relationships between principals' scores and estimated value-added (see table H7 in appendix H). Higher full scores and domain 1 and 2 scores based on rounded averages were associated with higher estimated value-added in all subjects combined, at a marginally significant level. Rounded domain 4 scores were significantly positively related to estimated value-added in all subjects combined. Similarly, full scores and domain 1, 2, and 4 scores based on rounded averages were all significantly positively associated with estimated value-added in math.

Implications and limitations of the study

The distribution of scores may become even more compressed when scores become part of school leader evaluations if the high stakes incentivize lenient ratings. It will be important to continue to examine score variation to determine whether the differences in scores provide sufficiently meaningful information on performance differences that is supported by other evidence

The findings from the 2013/14 pilot described in this report indicate that the FFL has evidence of reliability and validity, both of which are desired components in an evaluation tool. A key strength of the FFL is its reliability, as measured by both internal consistency and year-to-year stability. The internal consistency of the full FFL is high: school leaders identified as effective or ineffective on one domain tended to be identified in a similar way on the other domains. The FFL also exhibits moderate year-to-year score stability, comparable to that of widely used teacher observation instruments. A principal's full score and domain scores in the first pilot year were partially predictive of the principal's full score and domain scores in the second pilot year. Although an individual principal's scores might vary somewhat from one year to the next, no wide fluctuations in scores were observed that would raise concerns about reliability.

Data from the second pilot year (2013/14) also provided the first tentative evidence of the FFL's concurrent validity. Scores are, to some extent, differentiating principals who make larger or smaller contributions to student achievement growth. Higher full scores and scores in two of the four domains are significantly or marginally significantly associated with value-added in all subjects combined and value-added in math specifically. This evidence of concurrent validity sets the FFL apart from other principal evaluation tools examined in the literature. Only two other studies have examined validity of principal evaluation tools, focusing on a small number of district-specific evaluation instruments. Neither study found any robust evidence of a relationship between the instruments and principals' value-added (Grissom et al., 2015; Milanowski & Kimball, 2012).

One area where additional examination of the FFL may be warranted, particularly during wider implementation, is the distribution of scores. In both the 2012/13 and 2013/14 pilot years most school leaders scored in the upper third of the rating scale despite an average estimated value-added that was not statistically distinguishable from the state average. This suggests that supervisors tend to rate school leaders too leniently. Moreover, when FFL scores were calculated from whole number domain scores, variation was further reduced in both pilot years. Scores were all of low stakes during the pilot years. The variation may become even more compressed when scores become part of school leader evaluations if the high stakes incentivize lenient ratings. It will be important to continue to examine score variation to determine whether the differences in scores provide sufficiently meaningful information on performance differences that is supported by other evidence.

Limitations of the study

This section identifies limitations to consider when interpreting the findings of the study.

The interim report from this study developed a new method for measuring school leaders' value-added (see appendix G). Previous studies that measured principals' value-added used methods that mistakenly attribute the effectiveness of entire schools to the effectiveness of the principal alone or that permit comparisons only among small numbers of principals (Branch et al., 2012; Chiang et al., forthcoming; Coelli & Green, 2012; Dhuey & Smith, 2012, forthcoming; Grissom et al., 2015; 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.

Moreover, a valid measure for estimating the value-added of longer serving principals in the face of limited longitudinal data remains outstanding. This study was able to estimate value-added only for principals with six or fewer years of tenure at a school; the evidence of the concurrent validity of the FFL is therefore restricted to recently hired principals. It is unclear whether the estimated contributions to student achievement growth of longer serving principals are related to their FFL scores.

Based on the prevalence of high scores across two pilot years, the study team suggests using corroborative evidence to check that supervisors are correctly applying the standard for FFL scoring

Suggestions for improving Framework for Leadership evaluations

Although the study findings suggest that the FFL is a promising tool for reliably and validly measuring principal performance, questions remain about whether supervisors are too lenient when assigning FFL scores and whether the FFL is an appropriate tool for measuring assistant principal performance. This section includes steps that the Pennsylvania Department of Education and districts who are implementers or potential implementers of the FFL may wish to consider with regards to these outstanding questions.

Obtain ratings of school leaders by other stakeholders to check the validity of scores assigned by supervisors. Based on the prevalence of high scores across two pilot years, the study team suggests using corroborative evidence to check that supervisors are correctly applying the standard for FFL scoring. This step is especially important during the early years of FFL implementation as supervisors become familiar with the tool. One approach is to gather anonymous ratings of school leaders by other knowledgeable individuals, such as teachers. Although these ratings could be used for informative purposes rather than evaluative purposes, this approach 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). Evidence from ratings by teachers could be used 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. This evidence could also yield an assessment of 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).

Explore the most appropriate set of FFL components for measuring assistant principal performance. This study found that in both pilot years the internal consistency of the FFL for assistant principals was substantially lower than the internal consistency for

principals. Only one domain (domain 1) meets an acceptable level of internal consistency for assistant principals, and one of the three remaining domains (domain 4) does not meet the threshold even for a marginally acceptable level. This finding, taken with the lack of evidence of the validity of the FFL for assistant principals, suggests the need to continue to gather evidence on the statistical properties of the FFL for assistant principals specifically. The current role of the assistant principal in practice may not fit the construct of school leadership defined by some components of the FFL. The Pennsylvania Department of Education and other states or districts interested in implementing the FFL might consider either tailoring the FFL more specifically to the assistant principal position or redefining the role of the assistant principal in line with the school leader role conceptualized by the FFL. The latter would necessitate a more long-term approach.

Gather more evidence on the statistical properties of the FFL. In addition to these specific steps, it will be important to continue gathering evidence on the statistical properties of the FFL as the instrument is implemented across Pennsylvania. Monitoring wider implementation will confirm whether the FFL is a valid and reliable measure of performance across all school leaders in the state, not just among the sample of pilot participants. Also, continuing to gather evidence will enable the Pennsylvania Department of Education to examine additional measures of FFL validity and reliability and refine the FFL as needed. These measures might include, among others, convergent validity, as described earlier, and concurrent validity with measures of effectiveness other than value-added; for example, graduation and dropout rates using enrollment data or student safety and engagement using student surveys.

Continuing to gather evidence will enable the Pennsylvania Department of Education to examine additional measures of FFL validity and reliability and refine the FFL as needed

Appendix A. Prior research on measuring principal effectiveness

The properties 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. The interim report from this study (Teh et al., 2014) provided one of the few existing in-depth analyses of the reliability, score variation, and concurrent validity of a school leader evaluation tool intended for widespread use. With the few exceptions described below, the available statistical information on most other school leader 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.

Estimating principals' contributions to student achievement growth is essential for assessing whether evaluation tools accurately distinguish principals with larger and smaller contributions to student achievement growth. Only a few studies have developed and analyzed methods for estimating principals' contributions to student achievement growth (Branch et al., 2012; Chiang et al., forthcoming; Coelli & Green, 2012; Dhuey & Smith, 2012, forthcoming; Grissom et al., 2015; 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 achievement 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., forthcoming). 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 G).

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 three studies have examined the relationship between principal evaluation tools and value-added. Two of those studies focused on a small number of districts. In a study of 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); the study did not examine the relationship with the principals' own value-added. In Miami-Dade County Public Schools, principals' scores were positively associated with the value-added of their schools but not with the value-added of the principals themselves (Grissom et al., 2015). To date, the interim report from the current study (Teh et al., 2014) provides the only existing analysis of relationships between principals' value-added and scores from a school leader evaluation tool intended for statewide implementation. That report found no relationship between principals' value-added and scores on Pennsylvania's Framework for Leadership in the 2012/13 pilot year of the evaluation tool.

Developers of some principal evaluation tools have assessed their validity through approaches other than examining relationships with principals' 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 modest positive correlations in the range of .13–.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 .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 20 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.

In the 2012/13 pilot year the FFL included 19 components. In the 2013/14 pilot year an additional component was added in domain 2 (systems leadership): ensures a high-quality, high-performing staff (2g). However, scores for the newly added component in the 2013/14 pilot year were not collected. As such, analysis for this study is specific to the 19 other components that were consistent across the 2012/13 and 2013/14 pilot years.

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 in order 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.

(continued)

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

Name of component	Description of component
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.
2g. Ensures a high-quality, high-performing staff	The school leader establishes, supports and effectively manages processes and systems that ensure a high-quality, high-performing staff.
3: Leadership for learning	
3a. Leads school improvement initiatives	The school leader develops, implements, 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.
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 professional responsibilities through parent involvement and community engagement	The school leader designs structures and processes that result in parent involvement and community engagement, as well as support and ownership for the school.
4b. Shows professionalism	The school 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 2013/14 pilot year: Participants, evaluation procedures, and available data

Participants. All the FFL scores used in this report came from the 2013/14 pilot year. Understanding the criteria for participation in the 2013/14 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 541 schools spread across 193 local education agencies (table C1). The study's analyses included 640 school leaders—517 principals and 123 assistant principals—with scores from the 2013/14 pilot year. Collectively, these leaders were rated by 237 supervisors, 104 of whom had also rated at least one school leader participant in the 2012/13 pilot year.

Similar to participation in 2012/13 pilot year, local education agencies and schools that participated in the 2013/14 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 (152 of 193) were required to participate because they received Race to the Top funds (table C2). Most of the principals (401 of 517) and assistant principals (106 of 123) in the study were leaders in these 152 local education agencies.

Characteristics of students enrolled in schools that did and did not participate in the 2013/14 pilot year are shown in table C3; characteristics of participating and nonparticipating school leaders are shown in table C4.

Evaluation procedures. Similar to the 2012/13 pilot year, one supervising administrator evaluated each school leader in the 2013/14 pilot year. Superintendents and assistant superintendents constituted the majority of supervisors who rated principals (75 percent) and

Table C1. Number of participants in the 2013/14 Framework for Leadership pilot year

Type of participant	Count
Local education agencies (districts, charter schools, technical centers)	193
Schools	541
School leaders who received ratings	640
Principals	517
Assistant principals	123
Supervisors who assigned ratings	237

Source: Authors' calculations based on Framework for Leadership 2013/14 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 2013/14 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)	136	317	84
Receives Race to the Top Funds and has school receiving School Improvement Grant funds for transformation	16	84	22
Has school receiving School Improvement Grant funds for transformation (and no other reason)	1	3	2
Volunteer	38	109	15
Reason not recorded	2	4	0

Source: Authors' calculations based on Framework for Leadership 2013/14 pilot evaluation scores provided by the Pennsylvania Department of Education.

Table C3. Characteristics of students in Pennsylvania in 2013/14, by whether their school participated in the Framework for Leadership 2013/14 pilot year (percent unless otherwise indicated)

	Grades 4–5		Grades 6–8		Grades 9–12	
Student characteristic	Statewide	2013/14 pilot schools	Statewide	2013/14 pilot schools	Statewide	2013/14 pilot schools
Number of students	247,286	45,288	375,847	71,311	282,629	54,381
Baseline math score[a] (average *z*-score)	0.02	–0.03	0.02	–0.02	–0.05	–0.08
Baseline reading score[a] (average *z*-score)	0.02	–0.04	0.02	–0.02	–0.04	–0.08
Receives free lunch	40.2	42.4	37.7	38.5	34.4	35.3
Receives reduced-price lunch	5.0	5.7	5.2	5.9	5.4	6.2
English learner student	2.3	1.8	2.1	1.3	1.7	1.3
Has any disability	18.4	18.9	17.3	17.9	16.6	16.3
Moved schools during school year	3.5	3.7	3.7	4.3	11.7	10.6
Grade repeater	0.2	0.2	0.7	0.9	4.2	4.4
Over age for grade	0.2	0.2	0.3	0.3	0.9	0.9
Age (average years)	10.1	10.1	12.6	12.7	15.6	15.6
Female	49.2	48.8	49.0	48.4	49.3	49.4
Race/ethnicity						
Asian or Pacific Islander	3.7	2.6	3.4	2.1	2.8	1.9
Black, non-Hispanic	14.3	14.4	14.1	12.8	14.1	12.4
Hispanic	9.3	9.9	8.3	7.4	7.3	8.0
White, non-Hispanic	69.3	70.0	70.6	73.9	71.2	73.0
Other	2.7	2.4	1.3	1.5	0.6	0.5

Note: Based only on students who were included in at least one value-added model described in appendix F.

a. 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 in 2013/14, by whether they participated in the Framework for Leadership 2013/14 school year (percent unless otherwise indicated)

Characteristic	Principals		Assistant principals	
	Statewide	Participated in pilot	Statewide	Participated in pilot
Highest degree attained				
Bachelor's	14.7	11.9	11.1	9.4
Master's	75.7	79.1	85.4	85.8
Doctorate	9.6	9.0	3.5	4.7
Total experience in PK–12 education (average years)	18.2	18.1	14.7	13.7
Race/ethnicity				
White, non-Hispanic	87.7	90.6	86.5	95.3
Black, non-Hispanic	10.3	8.8	11.0	3.9
Other	2.0	0.6	2.5	0.8
Gender				
Female	44.5	39.7	40.8	37.8
Male	55.5	60.3	59.2	62.2

PK–12 is prekindergarten to grade 12.

Note: Percentages may not sum to 100 because of rounding.

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

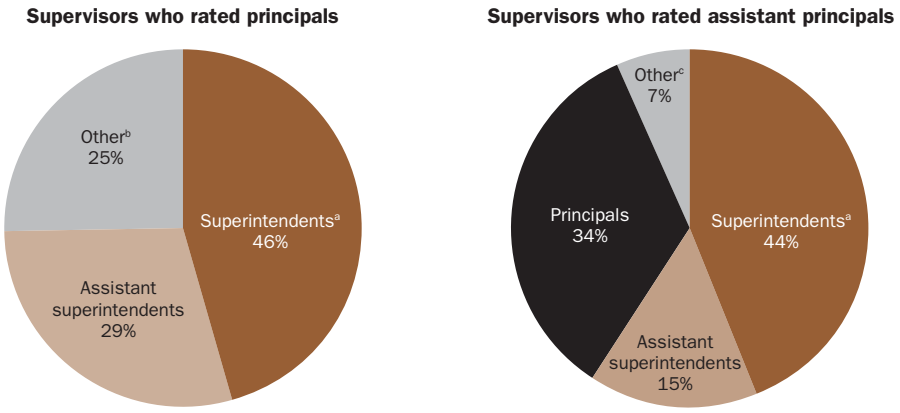
who rated assistant principals (59 percent; figure C1). The remaining quarter of supervisors who rated principals included directors of vocational education, other principals, supervisors of curriculum and instruction, supervisors of elementary education, and supervisors of secondary education. A third of the supervisors who rated assistant principals were the principals to whom the assistant principals were accountable.

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 and 2013/14 pilot years 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. Intermediate unit leaders who had previously participated in the training for the 2012/13 pilot year participated in a "refresher" train-the-trainer program for the 2013/14 pilot year. 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.
- The connectedness between the FFL and the Danielson Framework for Teaching.
- 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

Figure C1. Most supervisors in the Framework for Leadership 2013/14 pilot year were superintendents or assistant superintendents



a. Includes charter school chief executive officers.

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

c. Includes supervisors of curriculum and instruction.

Source: Authors' calculations based on Framework for Leadership 2013/14 pilot evaluation scores provided by the Pennsylvania Department of Education.

covered topics similar to those in the train-the-trainer session. For the 2013/14 pilot year only, the training sessions discussed concrete examples of the evidence that would merit a proficient score for every FFL component. Participants also received handouts on suggested questions to use to guide strategic discussions between supervisors and principals and between principals and teachers.

Participants in the 2013/2014 pilot year had some discretion over which components of the FFL would be included in the pilot evaluations. However, each pilot evaluation was supposed to include at least two components from each of the four 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 mid-year 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. The directive from the Pennsylvania Department of Instruction to include at least two components per domain in the evaluation was implemented with high fidelity. The 640 school leaders in the analysis (see table C1) were evaluated on at least two components from every domain; they constitute 100 percent of the principals and assistant principals who had a score from any component of the FFL. Moreover, the 640 school leaders in the analysis were typically evaluated on most of the components in the FFL; their pilot evaluations used an average of 16 out of 19 components, and 69 percent of the evaluations used all components. The average number of components

used did not change from the 2012/13 pilot year to the 2013/14 pilot year; the frequency of using all components was also similar across pilot years.

Since 2014/15 FFL evaluations have required supervisors to assign a domain score based on the preponderance of evidence within a domain, but supervisors assigned only component scores in the 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 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 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 2013/14. 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 2013/14; 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 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 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. The methods used for this report are identical to those used for the interim report (Teh et al., 2014).

Detailed tabulations of component score distributions

On every FFL component, the large majority of school leaders received a score of either proficient or distinguished (see figures 1 and 2 in the main report). Detailed tabulations of the percentages of principals (table D1) and assistant principals (table D2) receiving each of the four possible scores highlight the rarity of needs improvement and failing ratings.

Formal analysis of component difficulty

When school leaders and their supervisors determine which components to use in an evaluation, the selection process may compromise the fairness of evaluation scores. If there are

Table D1. Summary statistics on the distribution of Framework for Leadership component scores for principals, 2013/14

Component	Percentage of principals receiving: Failing (0 points)	Needs improvement (1 point)	Proficient (2 points)	Distinguished (3 points)	Mean score	Standard deviation
1a: Strategic goals	0.5	6.1	72.6	20.8	2.1	0.5
1b: Data for decisionmaking	0.0	7.7	66.0	26.3	2.2	0.6
1c: Empowering work environment	0.2	6.6	63.9	29.3	2.2	0.6
1d: Continuous improvement	0.0	5.9	68.6	25.5	2.2	0.5
1e: Lessons from accomplishments and failures	0.0	2.5	71.6	25.9	2.2	0.5
2a: Leverages resources	0.2	4.2	76.8	18.7	2.1	0.5
2b: School safety	0.2	2.4	72.0	25.3	2.2	0.5
2c: Complies with mandates	0.3	3.3	80.4	16.1	2.1	0.4
2d: Clear expectations for students and staff	0.0	3.4	74.9	21.7	2.2	0.5
2e: Communicates effectively	0.2	4.5	72.2	23.0	2.2	0.5
2f: Manages conflict	0.2	7.1	75.3	17.4	2.1	0.5
3a: School improvement initiatives	0.2	7.8	73.6	18.3	2.1	0.5
3b: Aligns curricula and instruction	0.7	8.8	69.3	21.2	2.1	0.6
3c: High-quality instruction	0.4	7.4	73.1	19.2	2.1	0.5
3d: High expectations for students	0.2	5.1	72.2	22.4	2.2	0.5
3e: Maximizes instructional time	0.2	1.5	72.7	25.6	2.2	0.5
4a: Parent and community involvement	0.2	6.8	65.2	27.7	2.2	0.6
4b: Professionalism	0.7	2.1	58.9	38.3	2.3	0.6
4c: Supports professional growth	0.4	2.8	69.8	27.1	2.2	0.5
All components	0.3	5.1	70.9	23.8	2.2	0.5

Source: Authors' calculations based on Framework for Leadership pilot evaluation scores from 2013/14 provided by the Pennsylvania Department of Education.

Table D2. Summary statistics on the distribution of Framework for Leadership component scores for assistant principals, 2013/14

Component	Percentage of assistant principals receiving: Failing (0 points)	Needs improvement (1 point)	Proficient (2 points)	Distinguished (3 points)	Mean score	Standard deviation
1a: Strategic goals	0.0	2.0	85.7	12.2	2.1	0.4
1b: Data for decisionmaking	0.0	7.0	77.4	15.7	2.1	0.5
1c: Empowering work environment	0.0	1.8	78.4	19.8	2.2	0.4
1d: Continuous improvement	0.0	3.3	72.8	23.9	2.2	0.5
1e: Lessons from accomplishments and failures	0.0	1.1	79.1	19.8	2.2	0.4
2a: Leverages resources	0.0	8.3	86.9	4.8	2.0	0.4
2b: School safety	0.0	0.0	76.3	23.7	2.2	0.4
2c: Complies with mandates	0.0	3.4	85.1	11.5	2.1	0.4
2d: Clear expectations for students and staff	0.9	1.8	77.2	20.2	2.2	0.5
2e: Communicates effectively	0.0	2.7	84.7	12.6	2.1	0.4
2f: Manages conflict	0.0	8.7	71.7	19.6	2.1	0.5
3a: School improvement initiatives	0.0	5.4	81.7	12.9	2.1	0.4
3b: Aligns curricula and instruction	0.0	12.1	78.0	9.9	2.0	0.5
3c: High-quality instruction	0.9	2.7	83.8	12.6	2.1	0.4
3d: High expectations for students	0.0	0.9	80.0	19.1	2.2	0.4
3e: Maximizes instructional time	0.0	3.3	85.7	11.0	2.1	0.4
4a: Parent and community involvement	0.0	3.4	72.9	23.7	2.2	0.5
4b: Professionalism	0.0	0.0	57.4	42.6	2.4	0.5
4c: Supports professional growth	0.0	1.8	78.6	19.6	2.2	0.4
All components	0.1	3.5	78.5	17.9	2.1	0.4

Source: Authors' calculations based on Framework for Leadership pilot evaluation scores from 2013/14 provided by the Pennsylvania Department of Education.

differences in relative difficulty of the components, school leaders may have an incentive to choose components that are substantially easier to score well in. Assessing the difficulty of each component can help determine whether concerns about fairness are substantiated.

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 2.3 for principals (see table D1) and from 2.0 to 2.4 for assistant principals (see table D2). 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, within each domain, school leaders and their supervisors could choose which two (or more) 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.

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* received by school leader *i*, the following regression was estimated:

(D1) $$y_{ci} = \alpha_c + \vartheta_i + \in_{ci},$$

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.

Adjusted average scores on the components (table D3) closely mirror the unadjusted average scores and thus confirm the conclusion drawn from the unadjusted averages: components were generally similar in difficulty. Adjusted average scores ranged from 2.1 to 2.3 for principals and from 2.0 to 2.4 for assistant principals. Both the unadjusted and adjusted average score ranges for principals and assistant principals were similar in the 2012/13 pilot year (Teh et al., 2014).

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

Because in 2013/14 most school leaders received component scores of proficient (2 points) or distinguished (3 points), which was consistent with the component score distribution in

Table D3. Average Framework for Leadership component scores, adjusted for differences in the mix of school leaders evaluated on different components, 2013/14

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

Source: Authors' calculations based on Framework for Leadership pilot evaluation scores from 2013/14 provided by the Pennsylvania Department of Education.

2012/13, the domain scores and full FFL scores were again concentrated primarily in the range of 2–3 points. Histograms of full FFL scores (see figures 3 and 4 in the main text) show evidence that few school leaders scored below 2. Detailed tabulations substantiate the visual evidence from the histograms (tables D4 and D5).

Table D4. Distribution of principals' scores on the full Framework for Leadership and its domains, 2013/14 (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	2.1	2.1	2.3
Standard deviation of scores	0.3	0.4	0.3	0.4	0.4
Distribution of scores					
Below 0.5	0.0	0.0	0.2	0.4	0.0
At least 0.5, below 1.0	0.4	0.0	0.2	0.0	0.2
At least 1.0, below 1.5	1.4	2.7	1.0	3.7	1.9
At least 1.5, below 2.0	16.8	12.0	9.1	10.6	5.4
Exactly 2.0	18.8	37.9	44.7	45.1	44.9
Above 2.0, below 2.5	45.8	24.6	26.5	22.4	17.2
At least 2.5, below 3.0	14.5	15.1	14.7	11.6	18.4
Exactly 3.0	2.3	7.7	3.7	6.2	12.0
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	0.2	0.4	0.0
At least 0.5, below 1.0	0.4	0.0	0.0	0.0	0.0
At least 1.0, below 1.5	1.0	2.7	1.2	3.7	2.1
At least 1.5, below 2.0	4.3	0.0	0.0	0.0	0.0
Exactly 2.0	54.7	74.5	80.3	78.1	67.5
Above 2.0, below 2.5	15.1	0.0	0.0	0.0	0.0
At least 2.5, below 3.0	15.1	0.0	0.0	0.0	0.0
Exactly 3.0	9.5	22.8	18.4	17.8	30.4

FFL is Framework for Leadership.

Note: Percentages may not sum to 100 because of rounding.

Source: Authors' calculations based on Framework for Leadership pilot evaluation scores from 2013/14 provided by the Pennsylvania Department of Education.

Table D5. Distribution of assistant principals' scores on the full Framework for Leadership and its domains, 2013/14 (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	2.3
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	0.8	0.0	0.0
At least 1.0, below 1.5	0.8	0.8	0.8	2.4	0.0
At least 1.5, below 2.0	11.4	8.9	8.9	8.9	4.9
Exactly 2.0	24.4	50.4	55.3	59.3	43.9
Above 2.0, below 2.5	52.8	25.2	17.9	17.1	18.7
At least 2.5, below 3.0	9.8	11.4	13.8	8.9	26.0
Exactly 3.0	0.8	3.3	2.4	3.3	6.5
Based on domain scores rounded to whole numbers					
Average score	2.2	2.1	2.1	2.1	2.3
Standard deviation of scores	0.3	0.4	0.4	0.4	0.5
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	0.0	0.8	1.6	2.4	0.0
At least 1.5, below 2.0	2.4	0.0	0.0	0.0	0.0
Exactly 2.0	57.7	84.6	82.1	85.4	67.5
Above 2.0, below 2.5	19.5	0.0	0.0	0.0	0.0
At least 2.5, below 3.0	15.4	0.0	0.0	0.0	0.0
Exactly 3.0	4.9	14.6	16.3	12.2	32.5

FFL is Framework for Leadership.

Note: Percentages may not sum to 100 because of rounding.

Source: Authors' calculations based on Framework for Leadership pilot evaluation scores from 2013/14 provided by the Pennsylvania Department of Education.

Appendix E. 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 alpha to assess the internal consistency of a scale with k items is (Cronbach, 1951):

(E1) $$\alpha = \frac{k\bar{c}}{\bar{v} + k\text{–}1\ \bar{c}},$$

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 alpha 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 2013/14. In actual evaluations the Pennsylvania Department of Education instructs supervisors to use the preponderance of evidence from the components in each domain to determine the domain scores.

Cronbach's alpha for a specific domain was obtained by treating the components within the domain as the items in applying equation E1. For each domain the calculation is based on school leaders with scores on all components in the domain because the calculation of Cronbach's alpha relies on having complete data.

Supplementary findings on the internal consistency of the Framework for Leadership

Calculating Cronbach's alpha 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 Cronbach's alpha values are appreciably lower than the Cronbach's alpha for the full index, the excluded piece is contributing positively to internal consistency. If the resulting Cronbach's alpha values are appreciably higher than the Cronbach's alpha for the full index, the excluded piece is contributing negatively to internal consistency. The Cronbach's alpha values obtained by excluding particular domains and components are provided in tables E1 and E2.

Table E1. Cronbach's alpha values for the full Framework for Leadership scores in the 2013/14 pilot year when particular domains are excluded

Portion of the Framework for Leadership	Cronbach's alpha	
	Principals	Assistant principals
Full Framework for Leadership with all four domains	.90	.79
Framework for Leadership, excluding:		
Domain 1: Strategic/cultural leadership	.86	.71
Domain 2: Systems leadership	.88	.75
Domain 3: Leadership for learning	.86	.72
Domain 4: Professional and community leadership	.89	.77

Source: Authors' calculations based on Framework for Leadership 2013/14 pilot evaluation scores provided by the Pennsylvania Department of Education.

Table E2. Cronbach's alpha values for Framework for Leadership domains in the 2013/14 pilot year when particular components are excluded

Domain and component	Cronbach's alpha	
	Principals	Assistant principals
Domain 1: Strategic/cultural leadership, excluding:		
No components	.77	.73
1a: Strategic goals	.70	.66
1b: Data for decisionmaking	.74	.65
1c: Empowering work environment	.74	.68
1d: Continuous improvement	.72	.63
1e: Lessons from accomplishments and failures	.75	.76
Domain 2: Systems leadership, excluding:		
No components	.75	.66
2a: Leverages resources	.71	.66
2b: School safety	.74	.62
2c: Complies with mandates	.70	.61
2d: Clear expectations for students and staff	.71	.57
2e: Communicates effectively	.71	.61
2f: Manages conflict	.73	.64
Domain 3: Leadership for learning, excluding:		
No components	.79	.67
3a: School improvement initiatives	.75	.59
3b: Aligns curricula and instruction	.74	.64
3c: High-quality instruction	.75	.61
3d: High expectations for students	.77	.65
3e: Maximizes instructional time	.77	.61
Domain 4: Professional and community leadership, excluding:		
No components	.61	.58
4a: Parent and community involvement	.60	.61
4b: Professionalism	.46	.30
4c: Supports professional growth	.46	.48

Source: Authors' calculations based on Framework for Leadership 2013/14 pilot evaluation scores provided by the Pennsylvania Department of Education.

Appendix F. Technical details and supplementary findings on year-to-year stability

This appendix provides technical details on how year-to-year score stability was calculated for the Framework for Leadership (FFL) and gives supplementary findings on year-to-year stability for subsamples and individual domains and components.

Calculating year-to-year stability for the Framework for Leadership

To measure the year-to-year stability of FFL scores, the study limited the sample of school leaders to those who were rated in both the 2012/13 and 2013/14 pilot years. Pearson's correlation coefficients were calculated for the full, domain, and component scores across years, separately for principals and assistant principals. For the sample of principals, correlations were also calculated for principals rated by the same supervisor in both pilot years, principals rated by a different supervisor in each year, principals rated on the same set of FFL components in both years, and principals rated on a different set of components in each year. Additional correlations of full and domain scores were calculated for three sufficiently large subsamples of the four main subgroups: principals rated by the same supervisor on the same set of components in both years, principals rated by the same supervisor on a different set of components in each year, and principals rated by a different supervisor on the same set of components in each year.

Full and domain scores for principals were moderately stable across years (tables F1 and F2). The full FFL score correlation across years was 0.54. Among the 184 principals participating in both pilot years, 23 percent had a total FFL score in 2013/14 that was within 0.05 point of their 2012/13 full score, 65 percent had a score within 0.25 point, and 89 percent had a score within 0.5 point. Full score stability was higher among principals rated by the same supervisor than among those rated by a different supervisor but still in the moderate range (.60). Even more so than rater consistency, the use of the same set of components for evaluations across years appears to play an important role in score stability. A majority of principals rated by the same supervisor were also rated on the same sets of components in

Table F1. Correlations of principals' 2012/13 and 2013/14 scores for the full Framework for Leadership and its domains, by rater or set of components rated

Portion of the Framework for Leadership	Year-to-year score correlation				
	All principals (*n*=184)	Principals with the same rater in both years (*n*=149)	Principals with a different rater in each year (*n*=35)	Principals rated on the same set of components in both years (*n*=127)	Principals rated on a different set of components in each year (*n*=57)
Full Framework for Leadership	.54*	.60*	.40*	.60*	.33*
Domain 1: Strategic/cultural Leadership	.47*	.50*	.42*	.53*	.27
Domain 2: Systems leadership	.45*	.50*	.31	.49*	.34*
Domain 3: Leadership for learning	.49*	.53*	.38*	.56*	.18
Domain 4: Professional and community leadership	.41*	.44*	.29	.52*	.12

* Significant at $p < .05$.

Source: Authors' calculations based on Framework for Leadership pilot evaluation scores from 2012/13 and 2013/14 provided by the Pennsylvania Department of Education.

Table F2. Correlations of principals' 2012/13 and 2013/14 scores for the full Framework for Leadership and its domains, by rater and set of components rated

Portion of the Framework for Leadership	Year-to-year score correlation				
	All principals (*n*=184)	Principals with the same rater and same set of components in both years (*n*=99)	Principals with the same rater and a different set of components in each year (*n*=50)	Principals with a different rater and the same set of components in each year (*n*=28)	Principals with a different rater and a different set of components in each year (*n*=7)
Full Framework for Leadership	.54*	.68*	.38*	.42*	na
Domain 1: Strategic/cultural Leadership	.47*	.58*	.26	.42*	na
Domain 2: Systems leadership	.45*	.55*	.42*	.32	na
Domain 3: Leadership for learning	.49*	.63*	.19	.39*	na
Domain 4: Professional and community leadership	.41*	.58*	.16	.35	na

* Significant at $p < .05$.

na is not applicable.

Source: Authors' calculations based on Framework for Leadership pilot evaluation scores from 2012/13 and 2013/14 provided by the Pennsylvania Department of Education.

both years; scores for this subgroup were the most stable (.68). Among principals who were rated by the same supervisor in both years but were rated on a different set of components in each year, score stability was low (.38).

Full scores for assistant principals were highly stable across years (table F3). Three of four domain scores had year-to-year correlations exceeding moderate stability and approaching high stability. However, the sample of assistant principals was limited.

Table F3. Correlations of assistant principals' 2012/13 and 2013/14 scores for the full Framework for Leadership and its domains

Portion of the Framework for Leadership	Year-to-year score correlation				
	All assistant principals (*n*=26)	Assistant principals with the same rater in both years (*n*=19)	Assistant principals with a different rater in each year (*n*=7)	Assistant principals rated on the same set of components in both years (*n*=17)	Assistant principals rated on a different set of components in each year (*n*=9)
Full Framework for Leadership	.80*	na	na	na	na
Domain 1: Strategic/cultural Leadership	.63*	na	na	na	na
Domain 2: Systems leadership	.22	na	na	na	na
Domain 3: Leadership for learning	.69*	na	na	na	na
Domain 4: Professional and community leadership	.76*	na	na	na	na

* Significant at $p < .05$.

na is not applicable.

Note: Because of the limited number of assistant principals participating in both pilot years, correlations are reported for the full sample of assistant principals only and not for any subsamples.

Source: Authors' calculations based on Framework for Leadership pilot evaluation scores from 2012/13 and 2013/14 provided by the Pennsylvania Department of Education.

Table F4. Correlations of principals' 2012/13 and 2013/14 component scores

Component	Year-to-year score correlation				
	All principals (n=184)	Principals with the same rater in both years (n=149)	Principals with a different rater in each year (n=35)	Principals rated on the same set of components in both years (n=127)	Principals rated on a different set of components in each year (n=57)
Domain 1: Strategic/cultural leadership					
1a: Strategic goals	.45*	.47*	.41*	.49*	na
1b: Data for decisionmaking	.34*	.37*	.26	.38*	na
1c: Empowering work environment	.39*	.42*	.33	.40*	na
1d: Continuous improvement	.44*	.46*	.37*	.44*	na
1e: Lessons from accomplishments and failures	.20*	.16	.33	.21*	na
Domain 2: Systems leadership					
2a: Leverages resources	.37*	.39*	.23	.39*	na
2b: School safety	.34*	.34*	.35*	.37*	na
2c: Complies with mandates	.36*	.36*	.36*	.38*	na
2d: Clear expectations for students and staff	.26*	.32*	.08	.31*	na
2e: Communicates effectively	.28*	.29*	.23	.30*	na
2f: Manages conflict	.22*	.30*	–.04	.23*	na
Domain 3: Leadership for learning					
3a: School improvement initiatives	.51*	.56*	.34	.53*	na
3b: Aligns curricula and instruction	.45*	.55*	.07	.46*	na
3c: High-quality instruction	.39*	.48*	.09	.45*	na
3d: High expectations for students	.32*	.30*	.37*	.39*	na
3e: Maximizes instructional time	.35*	.33*	.37*	.38*	na
Domain 4: Professional and community leadership					
4a: Parent and community involvement	.37*	.39*	.32	.42*	na
4b: Professionalism	.46*	.53*	.24	.46*	na
4c: Supports professional growth	.25*	.25*	.22	.24*	na

* Significant at $p < .05$.

na is not applicable.

Source: Authors' calculations based on Framework for Leadership pilot evaluation scores from 2012/13 and 2013/14 provided by the Pennsylvania Department of Education.

Table F5. Correlations of assistant principals' 2012/13 and 2013/14 component scores

Component	Year-to-year score correlation for all assistant principals (*n*=26)
Domain 1: Strategic/cultural leadership	
1a: Strategic goals	.31
1b: Data for decisionmaking	.33
1c: Empowering work environment	.06
1d: Continuous improvement	.70*
1e: Lessons from accomplishments and failures	.22
Domain 2: Systems leadership	
2a: Leverages resources	.50*
2b: School safety	–.09
2c: Complies with mandates	–.14
2d: Clear expectations for students and staff	.38
2e: Communicates effectively	.63*
2f: Manages conflict	.41
Domain 3: Leadership for learning	
3a: School improvement initiatives	.23
3b: Aligns curricula and instruction	.71*
3c: High-quality instruction	.55*
3d: High expectations for students	.21
3e: Maximizes instructional time	.32
Domain 4: Professional and community leadership	
4a: Parent and community involvement	.21
4b: Professionalism	.54*
4c: Supports professional growth	.40

* Significant at $p < .05$.

Source: Authors' calculations based on Framework for Leadership pilot evaluation scores from 2012/13 and 2013/14 provided by the Pennsylvania Department of Education.

Appendix G. Technical details of school and principal value-added models

In this study, principals' contributions to student achievement growth were estimated using value-added models (VAMs) for recently hired principals (those who began their current positions in 2008/09 or later). These contributions were therefore referred to as the principals' value-added. The starting point for estimating principals' 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 principals' contribution from the influences of other school-specific factors. This appendix provides details of the estimation of both school value-added and principals' value-added, which follows the same approach as in the interim report (Teh et al., 2014).

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:[5]

- PSSA math: grades 4–8 (2007/08–2013/14) and 11 (2009/10–2011/12).
- PSSA reading: grades 4–8 (2007/08–2013/14) and 11 (2009/10–2011/12).
- PSSA science: grades 4 and 8 (2007/08–2013/14) and 11 (2009/10–2011/12).
- PSSA writing: grades 5 and 8 (2007/08–2013/14) and 11 (2009/10–2011/12).
- Keystone algebra I, English literature, and biology: all spring scores for students in grade 8 or higher (2012/13–2013/14).

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

(G1) $$A_{isy} = \beta' \boldsymbol{P}_{i(y-1)} + \gamma' \boldsymbol{X}_{iy} + \delta' \boldsymbol{S}_{isy} + e_{isy}.$$

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 $\boldsymbol{P}_{i(y-1)}$ included variables for student i's prior-year PSSA scores. All the VAMs described by equation G1 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-year scores for grade repeaters came from the same grade as the outcome variable. The vector $\boldsymbol{P}_{i(y-1)}$ therefore included separate sets of variables for the prior-year scores of grade nonrepeaters and grade repeaters. The vector $\boldsymbol{X}_{iy}$ 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 $\boldsymbol{S}_{isy}$ included a school indicator 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 $\boldsymbol{S}_{isy}$. 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 G1, 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-year achievement scores as indicators of their academic abilities, but standardized tests are imperfect measures of ability. The measurement error introduced by using prior-year 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 G1 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 G1 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:

(G2a) $$\hat{A}_{isy} = A_{isy} - \beta' P_{i(y-1)}$$ [for students in grades 4–8]

(G2b) $$\hat{A}_{isy} = A_{isy} - \beta' P_{i(grade\ 8)}$$ [for students in grades 9–12]

The second step was to use the adjusted outcome in place of the actual score and estimate equation G3 by ordinary least squares separately for each grade-subject-year or assessment-year combination:

(G3) $$\hat{A}_{isy} = \gamma' \mathbf{X}_{iy} + \delta' \mathbf{S}_{isy} + e_{isy}.$$

The standard errors for the estimates from equation G3 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 high school students. Students who repeated a grade were included in the VAMs.[6] 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 the 2013/14 pilot year for students who did not repeat a grade are shown in table G1. 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.

Table G1. Assessments used as outcomes and baselines in the school value-added models, 2013/14

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. No PSSAs were administered in grade 11 in 2013/14.

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

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.[7] The imputations were 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 performance and outside the control of schools (table G2). 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. As in the analysis of Walsh & Lipscomb (2013), the gender and race/ethnicity controls were not meant to set different standards for students but to recognize that these variables explained statistically significant portions of the variation in student performance even after accounting for students' prior-year test scores and the other factors shown in table G2. 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.

Table G2. Student background control variables used in the school value-added models, 2013/14

Student background control variable	Definition
Free lunch	Free lunch participation (0 or 1)
Reduced-price lunch	Reduced-price lunch participation (0 or 1)
English learner student	English 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.

The sample characteristics of the school VAMs for 2013/14 are shown in table G3. The first column of data shows the error-adjusted standard deviation of school value-added—a 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 0.17 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 0.17 student-level standard deviation, which is equivalent to lifting the median-performing student in the state to the 57th percentile of performance. 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 2013/14.

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 the value-added estimates across subjects for the school.
- A math composite.
- A reading and writing composite.
- A science composite.

Table G3. Sample characteristics of school value-added models, 2013/14

Outcome	Error-adjusted standard deviation of school value-added (student *z*-score units)	Number of students	Number of schools
PSSA math, grade 4	0.17	123,338	1,607
PSSA math, grade 5	0.17	122,975	1,495
PSSA math, grade 6	0.18	121,903	1,090
PSSA math, grade 7	0.16	125,705	881
PSSA math, grade 8	0.16	126,780	871
Keystone algebra I	0.37	220,788	1,236
PSSA reading, grade 4	0.14	123,026	1,607
PSSA reading, grade 5	0.13	122,647	1,495
PSSA reading, grade 6	0.12	121,595	1,089
PSSA reading, grade 7	0.12	125,367	881
PSSA reading, grade 8	0.10	126,405	872
Keystone literature	0.20	152,135	759
PSSA writing, grade 5	0.30	120,912	1,494
PSSA writing, grade 8	0.24	125,628	871
PSSA science, grade 4	0.20	123,174	1,607
PSSA science, grade 8	0.16	126,103	871
Keystone biology	0.24	175,406	768

PSSA is Pennsylvania System of School Assessment.

Note: No PSSAs were administered in grade 11 in 2013/14.

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

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.[8] 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 the schools were likely to be imprecise.

Estimating principal value-added for recently hired principals

Although models of principal effectiveness that compare each principal with other principals 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 principals in the 2013/14 school year. But 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., forthcoming). Therefore, the current study estimated principal value-added using the same approach as in the interim report by taking school value-added as the starting point and then, for recently hired principals (those starting their current positions in 2008/09 or later), making adjustments to account for the lingering influences of previous principals and other school-level factors.

To measure the value-added of recently hired principals, 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 principals started their current positions. Formally, for 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 principals and other persistent school-level factors beyond the current principals' 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 principals' predecessors if the current principals 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 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:

$$(G4) \qquad SVA_{ly} = \alpha_0 + \alpha_m MSVA_l + \alpha_r RSVA_l + \alpha_s SSVA_l + \alpha_h high_{ly} + \alpha_{hm}(high_{ly} * MSVA_l) + \alpha_{hr}(high_{ly} * RSVA_l) + \alpha_{hs}(high_{ly} * SSVA_l) + \Sigma_{y=9}^{12}\alpha_y Year_y + \varepsilon_{ly}.$$

For each principal, the residual from equation G4 was an estimate of his or her contribution to student achievement growth, adjusted for the effects of previous principals 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 principal. Estimated coefficients on the baseline school value-added measures from equation G4—shown separately for elementary/middle and high schools—are provided in table G4.[9] This model assumed that baseline school value-added fully captured the effects of the previous principal and all other school-specific factors beyond the current principal's control. It also assumed that the current principal'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 G4 was estimated separately for principals who had led their schools for one, two, three, four, five, and six years because the relationships between and baseline school value-added could have been different for principals with different tenure lengths.

The estimation samples included all principals 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 (2008/09–2013/14) 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 G4, only value-added estimates from 2013/14 for recently hired principals in the pilot were subsequently used to assess the concurrent validity of the FFL (see appendix H).

Because the measures of baseline school value-added in equation G4 were estimates, they had measurement error, which would bias the estimated coefficients on those variables toward 0 unless addressed. To account for measurement error, each baseline school value-added variable was adjusted by an empirical Bayes shrinkage procedure before being used in equation G4, such that the regression coefficient on the adjusted variable would no longer be attenuated. (This adjustment was made only for the baseline school value-added measures on the right side of the equation and not for the dependent variable.) 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.[10] 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

Table G4. Relationship between baseline and current school value-added for recently hired principals, using subject-specific composite value-added measures

Outcome subject	Tenure in current position (years)	Coefficient on baseline school value added						Number of principals
		Math		Reading/writing		Science		
		Elementary and middle schools	High schools	Elementary and middle schools	High schools	Elementary and middle schools	High schools	
All combined	1	0.06*	0.09	0.26**	0.44**	0.14**	0.19**	2,984
	2	0.04	0.17**	0.28**	0.18**	0.13**	0.18**	1,987
	3	–0.01	0.08	0.26**	0.17*	0.10**	0.02	1,282
	4	–0.07	0.17*	0.27**	0.12	0.07*	–0.00	811
	5	–0.01	na	0.29**	na	0.04	na	438
	6	–0.09	na	0.34**	na	0.05	na	205
Math	1	0.36**	0.46**	0.08*	0.20*	–0.01	0.05	2,983
	2	0.30**	0.32**	0.06	0.08	0.02	0.05	1,987
	3	0.26**	0.17*	0.10*	0.12	–0.02	–0.05	1,282
	4	0.15**	0.22*	0.11*	0.05	–0.03	–0.07	811
	5	0.19**	na	0.17*	na	–0.04	na	438
	6	0.07	na	0.21*	na	0.01	na	205
Reading/writing	1	–0.07*	–0.01	0.48**	0.74**	0.05*	0.07	2,983
	2	–0.07*	0.13	0.51**	0.37**	0.04	0.06	1,987
	3	–0.13**	–0.02	0.46**	0.38**	0.03	–0.11	1,282
	4	–0.16**	0.13	0.44**	0.33*	0.00	–0.08	811
	5	–0.03	na	0.44**	na	–0.03	na	437
	6	–0.12	na	0.51**	na	–0.03	na	205
Science	1	–0.17**	–0.13	–0.02	0.11	0.75**	0.67**	2,975
	2	–0.18**	0.10	0.07	–0.05	0.66**	0.61**	1,983
	3	–0.23**	0.05	0.07	–0.06	0.57**	0.33**	1,278
	4	–0.31**	0.11	0.14*	–0.13	0.49**	0.20**	809
	5	–0.30**	na	0.23**	na	0.45**	na	438
	6	–0.28*	na	0.12	na	0.39**	na	205

* 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 principals' job assignment data provided by the Pennsylvania Department of Education.

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 G4 that would have stemmed from such errors.

The study does not estimate value-added for assistant principals and for longer serving principals. For assistant principals, it is unclear how to separately identify their contributions from those of the principal. For longer serving principals (those who started their current positions before 2008/09), their baseline school value-added cannot be estimated using available data.

Longer serving principals could be included if they have had sufficient time to shape their school's value-added so that lingering effects of previous leaders and other school factors

were not relevant. That is, longer serving principals could be included if the effects of the school's baseline value-added on the school's current value-added were negligible. However, findings from the pilot study suggest that baseline school factors persist many years and that imposing the assumption that they do not matter is likely to produce biased value-added estimates for longer serving principals.

To see this clearly, the study estimated a variant of equation G4 for recently hired principals 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 G4, the model controlled for $high_{ly}$, an indicator of whether the principal 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 G4, the model allowed the relationship between baseline school value-added and current school value-added to be different for elementary/middle school principals and high school principals. The resulting regression equation had the following form:

$$(G5) \qquad SVA_{ly} = \alpha_0 + \alpha_1 CSVA_l + \alpha_h high_{ly} + \alpha_{hc}(high_{ly} * CSVA_l) + \sum_{y=9}^{12} \alpha_y Year_y + \varepsilon_{ly}.$$

To test the assumption that the lingering effects of previous principals would be negligible after the current principals had served for more than six years, equation G5 was estimated separately for principals who had led their schools for one, two, three, four, five, and six years. If the assumption were valid, α_1 and $(\alpha_1 + \alpha_{hc})$ would decrease monotonically with the current principal's length of service and approach zero. However, α_1 and $(\alpha_1 + \alpha_{hc})$ did not approach 0 (table G5). Therefore, the available measure of principal value-added for longer serving principals was less than ideal, and these principals were not included the analysis of the validity of the FFL.

Table G5. Relationship between baseline and current school value-added estimates for principals using composite value-added measures that combine all subjects

Principals who have led their current school for	Coefficient on baseline school value-added		Number of school leaders
	Elementary and middle schools	High schools	
1 year	0.46**	0.70**	3,042
2 years	0.44**	0.51**	2,016
3 years	0.35**	0.28**	1,296
4 years	0.26**	0.31**	824
5 years	0.32**	na	447
6 years	0.27**	na	209

** 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 principals' job assignment data provided by the Pennsylvania Department of Education.

The average value-added of principals in the pilot was similar to the average for all principals 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). 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 G6).

Table G6. Mean and standard deviation of the value-added estimates for recently hired principals participating in the Framework for Leadership 2013/14 pilot year relative to the statewide distribution of principals' value-added estimates

Outcome subject	Mean relative to statewide average (principal standard deviations)	Error-adjusted standard deviation (principal standard deviations)	Number of principals
All combined	0.05	0.94	305
Math	–0.01	0.96	305
Reading/writing	0.10	1.08	305
Science	0.08	0.94	305

Note: Recently hired principals are those who began their current positions in 2008/09 or later. No values were statistically significant.

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

Appendix H. Technical details and supplementary findings on the relationships between Framework for Leadership scores and principals' value-added

This appendix provides details on the method for estimating the relationship between the value-added of recently hired principals and their full Framework for Leadership (FFL) scores and some domain and component scores in 2013/14 and detailed results of the estimated relationships.

Estimation model

A regression model was used to estimate the relationships between the value-added of recently hired principals and the full, domain, and component FFL scores they received in 2013/14. 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 G. The regression model had the following form:

(H1) $$FFL_l = \beta_0 + \beta_1 VA_l + \boldsymbol{T}_l\boldsymbol{\delta} + \varepsilon_l$$

where $\boldsymbol{T}_l$ was a vector of five indicator variables identifying principals who had led their current school for two, three, four, five, and six years; ε_l was a random error term; and β_0, β_1, and $\boldsymbol{\delta}$ were coefficients that were estimated. The key coefficient of interest, β_1, measured the average change in the FFL score (measured in points on the FFL) for a unit change in principal value-added (measured in standard deviations of principal value-added). 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 principal value-added. The indicator variables for length of service in the current school accounted for the fact that value-added was estimated separately for—and was therefore not comparable across—principals with different lengths of service. This model was estimated only for principals in the 2013/14 pilot who began leading their current schools in 2008/09 or later.

The sample sizes in the 2013/14 pilot allowed for moderate levels of precision in estimating the relationship between principals' value-added and their FFL scores. Although the estimated relationships presented in this report are expressed as the regression coefficient (β_1) from equation H1, 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 2013/14 pilot, the study could reliably (with 80 percent power) detect a correlation between VA_l and FFL_l if the true correlation were at least .16. By comparison, prior research found a correlation of .24 between the Framework for Teaching and teachers' value-added in Pennsylvania (Walsh & Lipscomb, 2013). Therefore, the correlation between FFL scores and principals' value-added would be reliably detectable even if it was somewhat lower than the correlation between the Framework for Teaching and teachers' value-added.

Detailed results

The following tables contain detailed regression results of various versions of equation H1 where the dependent variable consisted of full, domain, or component FFL scores and the main independent variable was a principal's value-added in each of various subjects. In these tables, β_1 is expressed as the difference in scores between principals at the 84th and 50th percentiles of principal value-added. This is because a unit increase in principal value-added—an increase of one standard deviation of principal value-added—is equivalent to moving a principal previously at the 50th percentile to the 84th percentile of the value-added distribution. Tables H1–H7 present regression results where the dependent variable, the FFL score, and the main independent variable, a principal's value-added, are measured in the same school year. Similarly, tables H8–H12 present the correlation coefficients for the estimated correlations between a principal's FFL score and value-added, both measured in the 2013/14 pilot year. The correlation coefficients provide a measure of association that can be compared to similar correlations in other studies, including those in the Measures of Effective Teaching studies, and are adjusted for estimation error in the value-added measures.[11]

Some studies that have examined correlations between a teacher's classroom observation scores and value-added estimates use value-added estimates measured in a different year or with a different section of students from the classroom observation scores as a precaution (for example, see Kane and Staiger, 2012). This approach accounts for the possibility that there are external factors in any given year that are not observed and controlled for in a value-added model but still influence estimated contributions to student achievement growth and are also captured in observation scores. Using a value-added estimate from a different year ensures that this unmeasured factor (or error) in one year is not also correlated with the teacher observation score that is the dependent variable. As a sensitivity check, the study team also estimated cross-year correlations of FFL scores and value-added measures. Tables H13–H17 contain detailed regression results of versions of equation H1 in which the dependent variables, FFL scores, were measured in the 2013/14 pilot year and the independent variables, principals' value-added and length of service, were measured in the 2012/13 pilot year. Tables H18–H22 contain detailed regression results of versions of equation H1 in which the dependent variables, FFL scores, were measured in the 2012/13 pilot year and the independent variables, principals' value-added and length of service, were measured in the 2013/14 pilot year.

Table H1. Association between the Framework for Leadership scores in the 2013/14 pilot year and the value-added estimates for recently hired principals

Outcome	Value-added measure	Predicted difference in FFL score between principals at the 84th and 50th percentiles of value-added	
		Estimate	*p* value
Full Framework for Leadership score	All subjects	0.04	.070
	Math	0.05*	.017
	Reading/writing	0.02	.412
	Science	0.04	.060
Score on domain 1 (Strategic/cultural leadership)	All subjects	0.03	.242
	Math	0.05	.056
	Reading/writing	–0.00	.857
	Science	0.04	.138
Score on domain 2 (Systems leadership)	All subjects	0.05*	.037
	Math	0.05**	.008
	Reading/writing	0.02	.287
	Science	0.05*	.045
Score on domain 3 (Leadership for learning)	All subjects	0.03	.221
	Math	0.03	.214
	Reading/writing	0.02	.426
	Science	0.04	.121
Score on domain 4 (Professional and community leadership)	All subjects	0.06*	.048
	Math	0.07*	.011
	Reading/writing	0.03	.216
	Science	0.05	.089

* Significant at $p < .05$; ** significant at $p < .01$.

FFL is Framework for Leadership.

Note: Recently hired school leaders are those who began their current positions in 2008/09 or later; $n = 305$.

Source: Authors' calculations based on Framework for Leadership pilot evaluation data from 2013/14, student achievement and background data, and principals' job assignment data provided by the Pennsylvania Department of Education.

Table H2. Predicted difference in component scores between recently hired principals at the 84th and 50th percentiles of value-added in all subjects combined

Component	Estimate	*p* value
1a: Strategic goals	0.02	.512
1b: Data for decisionmaking	0.07	.057
1c: Empowering work environment	0.07	.085
1d: Continuous improvement	–0.04	.312
1e: Lessons from accomplishments and failures	0.01	.826
2a: Leverages resources	0.06	.081
2b: School safety	0.08*	.026
2c: Complies with mandates	0.00	.974
2d: Clear expectations for students and staff	–0.01	.734
2e: Communicates effectively	0.05	.163
2f: Manages conflict	0.05	.156
3a: School improvement initiatives	0.00	.974
3b: Aligns curricula and instruction	–0.03	.511
3c: High-quality instruction	0.06	.097
3d: High expectations for students	0.00	.979
3e: Maximizes instructional time	0.06	.086
4a: Parent and community involvement	0.02	.629
4b: Professionalism	0.09*	.020
4c: Supports professional growth	0.04	.242

* Significant at $p < .05$; ** significant at $p < .01$.

Note: Analyses are based on a value-added measure that combines all subjects. Recently hired principals are those who began their current positions in 2008/09 or later.

Source: Authors' calculations based on Framework for Leadership pilot evaluation data from 2013/14, student achievement and background data, and principals' job assignment data provided by the Pennsylvania Department of Education.

Table H3. Predicted difference in component scores between recently hired principals at the 84th and 50th percentiles of value-added in math

Component	Estimate	*p* value
1a: Strategic goals	0.03	.410
1b: Data for decisionmaking	0.08*	.029
1c: Empowering work environment	0.09*	.015
1d: Continuous improvement	–0.01	.838
1e: Lessons from accomplishments and failures	0.02	.616
2a: Leverages resources	0.08**	.004
2b: School safety	0.05	.111
2c: Complies with mandates	0.01	.684
2d: Clear expectations for students and staff	0.04	.137
2e: Communicates effectively	0.08*	.018
2f: Manages conflict	0.03	.410
3a: School improvement initiatives	0.01	.705
3b: Aligns curricula and instruction	0.00	.941
3c: High-quality instruction	0.06	.074
3d: High expectations for students	–0.03	.413
3e: Maximizes instructional time	0.05	.089
4a: Parent and community involvement	0.07	.055
4b: Professionalism	0.10*	.011
4c: Supports professional growth	0.03	.311

* Significant at $p < .05$; ** significant at $p < .01$.

Note: Analyses are based on a value-added measure that combines all subjects. Recently hired principals are those who began their current positions in 2008/09 or later.

Source: Authors' calculations based on Framework for Leadership pilot evaluation data from 2013/14, student achievement and background data, and principals' job assignment data provided by the Pennsylvania Department of Education.

Table H4. Predicted difference in component scores between recently hired principals at the 84th and 50th percentiles of value-added in reading and writing

Component	Estimate	*p* value
1a: Strategic goals	–0.01	.797
1b: Data for decisionmaking	0.02	.483
1c: Empowering work environment	0.01	.647
1d: Continuous improvement	–0.05	.111
1e: Lessons from accomplishments and failures	–0.02	.499
2a: Leverages resources	0.02	.540
2b: School safety	0.04	.179
2c: Complies with mandates	–0.01	.597
2d: Clear expectations for students and staff	–0.04	.086
2e: Communicates effectively	0.01	.704
2f: Manages conflict	0.04	.168
3a: School improvement initiatives	–0.01	.629
3b: Aligns curricula and instruction	–0.02	.565
3c: High-quality instruction	0.05	.154
3d: High expectations for students	0.01	.824
3e: Maximizes instructional time	0.03	.346
4a: Parent and community involvement	–0.02	.673
4b: Professionalism	0.06*	.049
4c: Supports professional growth	0.03	.350

* Significant at $p < .05$; ** significant at $p < .01$.

Note: Analyses are based on a value-added measure that combines all subjects. Recently hired principals are those who began their current positions in 2008/09 or later.

Source: Authors' calculations based on Framework for Leadership pilot evaluation data from 2013/14, student achievement and background data, and principals' job assignment data provided by the Pennsylvania Department of Education.

Table H5. Predicted difference in component scores between recently hired principals at the 84th and 50th percentiles of value-added in science

Component	Estimate	*p* value
1a: Strategic goals	0.04	.328
1b: Data for decisionmaking	0.06	.086
1c: Empowering work environment	0.04	.281
1d: Continuous improvement	–0.02	.588
1e: Lessons from accomplishments and failures	0.05	.214
2a: Leverages resources	0.02	.526
2b: School safety	0.11**	.002
2c: Complies with mandates	0.01	.738
2d: Clear expectations for students and staff	–0.00	.940
2e: Communicates effectively	0.01	.787
2f: Manages conflict	0.03	.476
3a: School improvement initiatives	0.03	.430
3b: Aligns curricula and instruction	–0.02	.550
3c: High-quality instruction	0.02	.564
3d: High expectations for students	0.05	.171
3e: Maximizes instructional time	0.04	.213
4a: Parent and community involvement	0.01	.781
4b: Professionalism	0.05	.240
4c: Supports professional growth	0.05	.152

* Significant at $p < .05$; ** significant at $p < .01$.

Note: Analyses are based on a value-added measure that combines all subjects. Recently hired principals are those who began their current positions in 2008/09 or later.

Source: Authors' calculations based on Framework for Leadership pilot evaluation data from 2013/14, student achievement and background data, and principals' job assignment data provided by the Pennsylvania Department of Education.

Table H6. Association between Framework for Leadership scores in the 2013/14 pilot year and value-added estimates for recently hired principals, by grade span

Grade span[a]	Outcome	Predicted difference in FFL score between principals at the 84th and 50th percentiles of value-added		Number of principals
		Estimate	*p* value	
Elementary	Full Framework for Leadership score	0.02	.487	155
	Score on domain 1 (Strategic/cultural leadership)	0.00	.928	155
	Score on domain 2 (Systems leadership)	0.03	.144	155
	Score on domain 3 (Leadership for learning)	0.00	.887	155
	Score on domain 4 (Professional and community leadership)	0.04	.349	155
Middle	Full Framework for Leadership score	0.12	.072	81
	Score on domain 1 (Strategic/cultural leadership)	0.14*	.040	81
	Score on domain 2 (Systems leadership)	0.12	.114	81
	Score on domain 3 (Leadership for learning)	0.12	.148	81
	Score on domain 4 (Professional and community leadership)	0.11	.096	81
High	Full Framework for Leadership score	–0.02	.627	69
	Score on domain 1 (Strategic/cultural leadership)	–0.04	.255	69
	Score on domain 2 (Systems leadership)	–0.01	.782	69
	Score on domain 3 (Leadership for learning)	–0.02	.518	69
	Score on domain 4 (Professional and community leadership)	0.01	.750	69

* Significant at $p < .05$; ** significant at $p < .01$.

FFL is 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 2013/14 pilot year who have a value-added measure. Recently hired principals are those who began their current positions in 2008/09 or later.

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 pilot evaluation data from 2013/14, student achievement and background data, and principals' job assignment data provided by the Pennsylvania Department of Education.

Table H7. Association between the full Framework for Leadership and domain scores in the 2013/14 pilot year and value-added estimates for recently hired principals, using rounded domain averages

Outcome	Value-added measure	Predicted difference in FFL score between principals at the 84th and 50th percentiles of value-added		Number of school leaders
		Estimate	*p* value	
Full Framework for Leadership score	All subjects	0.05	.061	305
	Math	0.06**	.007	305
	Reading/writing	0.02	.410	305
	Science	0.04	.135	305
Score on domain 1 (Strategic/cultural leadership)	All subjects	0.05	.098	305
	Math	0.07*	.013	305
	Reading/writing	0.01	.585	305
	Science	0.04	.184	305
Score on domain 2 (Systems leadership)	All subjects	0.05	.084	305
	Math	0.06*	.024	305
	Reading/writing	0.02	.414	305
	Science	0.05	.101	305
Score on domain 3 (Leadership for learning)	All subjects	0.02	.439	305
	Math	0.03	.330	305
	Reading/writing	0.01	.713	305
	Science	0.02	.418	305
Score on domain 4 (Professional and community leadership)	All subjects	0.07*	.038	305
	Math	0.09**	.002	305
	Reading/writing	0.03	.329	305
	Science	0.04	.216	305

* Significant at $p < .05$; ** significant at $p < .01$.

FFL is Framework for Leadership.

Note: Recently hired principals are those who began their current positions in 2008/09 or later. Domain scores are rounded averages, and the full FFL score is an average of the rounded domain scores.

Source: Authors' calculations based on Framework for Leadership pilot evaluation data from 2013/14, student achievement and background data, and principals' job assignment data provided by the Pennsylvania Department of Education.

Table H8. Correlations between the value-added of recently hired principals in the 2013/14 pilot year and their Framework for Leadership scores

Outcome	Value-added measure	Correlation	Number of school leaders
Full Framework for Leadership score	All subjects	.10	305
	Math	.11	305
	Reading/writing	.03	305
	Science	.12	305
Score on domain 1 (Strategic/cultural leadership)	All subjects	.05	305
	Math	.09	305
	Reading/writing	–.04	305
	Science	.08	305
Score on domain 2 (Systems leadership)	All subjects	.12*	305
	Math	.12*	305
	Reading/writing	.05	305
	Science	.13*	305
Score on domain 3 (Leadership for learning)	All subjects	.07	305
	Math	.06	305
	Reading/writing	.04	305
	Science	.10	305
Score on domain 4 (Professional and community leadership)	All subjects	.09	305
	Math	.11	305
	Reading/writing	.05	305
	Science	.09	305

* Significant at $p < .05$; ** significant at $p < .01$.

FFL is Framework for Leadership.

Note: Recently hired principals are those who began their current positions in 2008/09 or later. Correlations are adjusted for estimation error in value-added estimates (Jacob and Lefgren 2008).

Source: Authors' calculations based on Framework for Leadership pilot evaluation data from 2013/14, student achievement and background data, and principals job assignment data provided by the Pennsylvania Department of Education.

Table H9. Correlations between the value-added in all subjects combined of recently hired principals in the 2013/14 pilot year and their Framework for Leadership scores

Component	Correlation
1a: Strategic goals	.01
1b: Data for decisionmaking	.10
1c: Empowering work environment	.08
1d: Continuous improvement	–.08
1e: Lessons from accomplishments and failures	–.02
2a: Leverages resources	.10
2b: School safety	.14*
2c: Complies with mandates	–.02
2d: Clear expectations for students and staff	–.03
2e: Communicates effectively	.06
2f: Manages conflict	.09
3a: School improvement initiatives	–.02
3b: Aligns curricula and instruction	–.05
3c: High-quality instruction	.11
3d: High expectations for students	–.01
3e: Maximizes instructional time	.09
4a: Parent and community involvement	–.01
4b: Professionalism	.13*
4c: Supports professional growth	.06

* Significant at $p < .05$.

Note: Analyses are based on a value-added measure that combines all subjects. Recently hired principals are those who began their current positions in 2008/09 or later. Correlations are adjusted for estimation error in value-added estimates (Jacob & Lefgren, 2008).

Source: Authors' calculations based on Framework for Leadership pilot evaluation data from 2013/14, student achievement and background data, and principals' job assignment data provided by the Pennsylvania Department of Education.

Table H10. Correlations between the value-added in math of recently hired principals in the 2013/14 pilot year and their Framework for Leadership scores

Component	Correlation
1a: Strategic goals	.02
1b: Data for decisionmaking	.12
1c: Empowering work environment	.11
1d: Continuous improvement	–.02
1e: Lessons from accomplishments and failures	.00
2a: Leverages resources	.15*
2b: School safety	.08
2c: Complies with mandates	–.01
2d: Clear expectations for students and staff	.07
2e: Communicates effectively	.12
2f: Manages conflict	.05
3a: School improvement initiatives	–.01
3b: Aligns curricula and instruction	–.01
3c: High-quality instruction	.11
3d: High expectations for students	–.07
3e: Maximizes instructional time	.09
4a: Parent and community involvement	.08
4b: Professionalism	.13*
4c: Supports professional growth	.04

* Significant at $p < .05$; ** significant at $p < .01$.

Note: Analyses are based on a value-added measure in math. Recently hired principals are those who began their current positions in 2008/09 or later. Correlations are adjusted for estimation error in value-added estimates (Jacob and Lefgren 2008).

Source: Authors' calculations based on Framework for Leadership pilot evaluation data from 2013/14, student achievement and background data, and principals' job assignment data provided by the Pennsylvania Department of Education.

Table H11. Correlations between the value-added in reading and writing of recently hired principals in the 2013/14 pilot year and their Framework for Leadership scores

Component	Correlation
1a: Strategic goals	–.04
1b: Data for decisionmaking	.03
1c: Empowering work environment	.00
1d: Continuous improvement	–.12
1e: Lessons from accomplishments and failures	–.08
2a: Leverages resources	.03
2b: School safety	.08
2c: Complies with mandates	–.05
2d: Clear expectations for students and staff	–.11
2e: Communicates effectively	–.01
2f: Manages conflict	.09
3a: School improvement initiatives	–.04
3b: Aligns curricula and instruction	–.05
3c: High-quality instruction	.09
3d: High expectations for students	.00
3e: Maximizes instructional time	.04
4a: Parent and community involvement	–.07
4b: Professionalism	.11
4c: Supports professional growth	.05

Note: Analyses are based on a value-added measure that combines reading and writing. Recently hired principals are those who began their current positions in 2008/09 or later. Correlations are adjusted for estimation error in value-added estimates (Jacob and Lefgren 2008). No values were statistically significant.

Source: Authors' calculations based on Framework for Leadership pilot evaluation data from 2013/14, student achievement and background data, and principals' job assignment data provided by the Pennsylvania Department of Education.

Table H12. Correlations between the value-added in science of recently hired principals in the 2013/14 pilot year and their Framework for Leadership scores

Component	Correlation
1a: Strategic goals	.05
1b: Data for decisionmaking	.11
1c: Empowering work environment	.07
1d: Continuous improvement	–.05
1e: Lessons from accomplishments and failures	.07
2a: Leverages resources	.04
2b: School safety	.20**
2c: Complies with mandates	.01
2d: Clear expectations for students and staff	–.01
2e: Communicates effectively	.02
2f: Manages conflict	.06
3a: School improvement initiatives	.05
3b: Aligns curricula and instruction	–.05
3c: High-quality instruction	.05
3d: High expectations for students	.09
3e: Maximizes instructional time	.07
4a: Parent and community involvement	.00
4b: Professionalism	.06
4c: Supports professional growth	.09

* Significant at $p < .05$; ** significant at $p < .01$.

Note: Analyses are based on a value-added measure in science. Recently hired principals are those who began their current positions in 2008/09 or later. Correlations are adjusted for estimation error in value-added estimates (Jacob and Lefgren 2008).

Source: Authors' calculations based on Framework for Leadership pilot evaluation data from 2013/14, student achievement and background data, and principals' job assignment data provided by the Pennsylvania Department of Education.

Table H13. Association between the Framework for Leadership scores in the 2013/14 pilot year and the value-added estimates for recently hired principals in the 2012/13 pilot year

Outcome	Value-added measure	Predicted difference in FFL score between principals at the 84th and 50th percentiles of value-added		Number of school leaders
		Estimate	*p* value	
Full Framework for Leadership score	All subjects	0.04	.272	107
	Math	0.04	.325	107
	Reading/writing	0.04	.305	107
	Science	–0.01	.640	107
Score on domain 1 (Strategic/cultural leadership)	All subjects	0.09	.053	107
	Math	0.07	.125	107
	Reading/writing	0.07	.139	107
	Science	0.03	.387	107
Score on domain 2 (Systems leadership)	All subjects	0.02	.493	107
	Math	0.03	.519	107
	Reading/writing	0.04	.358	107
	Science	–0.04	.165	107
Score on domain 3 (Leadership for learning)	All subjects	0.04	.366	107
	Math	0.04	.404	107
	Reading/writing	0.05	.265	107
	Science	–0.02	.512	107
Score on domain 4 (Professional and community leadership)	All subjects	0.01	.789	107
	Math	0.02	.609	107
	Reading/writing	0.01	.823	107
	Science	–0.03	.548	107

FFL is Framework for Leadership.

Note: Recently hired principals are those who began their current positions in 2008/09 or later. No values were statistically significant.

Source: Authors' calculations based on Framework for Leadership pilot evaluation data from 2013/14, student achievement and background data, and principals' job assignment data provided by the Pennsylvania Department of Education.

Table H14. Predicted difference in component scores on the Framework for Leadership in the 2013/14 pilot year between recently hired principals at the 84th and 50th percentiles of value-added in all subjects combined in the 2012/13 pilot year

Component	Estimate	*p* value
1a: Strategic goals	0.06	.491
1b: Data for decisionmaking	0.19**	.002
1c: Empowering work environment	0.12	.073
1d: Continuous improvement	0.04	.587
1e: Lessons from accomplishments and failures	0.09	.140
2a: Leverages resources	0.07	.281
2b: School safety	–0.02	.784
2c: Complies with mandates	–0.05	.435
2d: Clear expectations for students and staff	–0.01	.921
2e: Communicates effectively	0.07	.404
2f: Manages conflict	0.13*	.030
3a: School improvement initiatives	0.06	.404
3b: Aligns curricula and instruction	0.01	.904
3c: High-quality instruction	0.10	.124
3d: High expectations for students	0.06	.310
3e: Maximizes instructional time	–0.01	.900
4a: Parent and community involvement	0.08	.228
4b: Professionalism	–0.01	.939
4c: Supports professional growth	–0.02	.812

* Significant at $p < .05$; ** significant at $p < .01$.

Note: Analyses are based on a value-added measure that combines all subjects. Recently hired principals are those who began their current positions in 2008/09 or later.

Source: Authors' calculations based on Framework for Leadership pilot evaluation data from 2013/14, student achievement and background data, and principals' job assignment data provided by the Pennsylvania Department of Education.

Table H15. Predicted difference in component scores on the Framework for Leadership in the 2013/14 pilot year between recently hired principals at the 84th and 50th percentiles of value-added in math in the 2012/13 pilot year

Component	Estimate	*p* value
1a: Strategic goals	0.04	.595
1b: Data for decisionmaking	0.18**	.003
1c: Empowering work environment	0.04	.466
1d: Continuous improvement	0.02	.738
1e: Lessons from accomplishments and failures	0.06	.248
2a: Leverages resources	0.11*	.044
2b: School safety	–0.02	.696
2c: Complies with mandates	–0.01	.915
2d: Clear expectations for students and staff	–0.01	.863
2e: Communicates effectively	0.05	.475
2f: Manages conflict	0.09	.192
3a: School improvement initiatives	0.10	.083
3b: Aligns curricula and instruction	–0.01	.900
3c: High-quality instruction	0.07	.307
3d: High expectations for students	0.03	.648
3e: Maximizes instructional time	0.04	.569
4a: Parent and community involvement	0.05	.388
4b: Professionalism	0.06	.429
4c: Supports professional growth	–0.02	.795

* Significant at $p < .05$; ** significant at $p < .01$.

Note: Analyses are based on a value-added measure that combines all subjects. Recently hired principals are those who began their current positions in 2008/09 or later.

Source: Authors' calculations based on Framework for Leadership pilot evaluation data from 2013/14, student achievement and background data, and principals' job assignment data provided by the Pennsylvania Department of Education.

Table H16. Predicted difference in component scores on the Framework for Leadership in the 2013/14 pilot year between recently hired principals at the 84th and 50th percentiles of value-added in reading and writing in the 2012/13 pilot year

Component	Estimate	*p* value
1a: Strategic goals	0.06	.467
1b: Data for decisionmaking	0.10	.099
1c: Empowering work environment	0.12	.095
1d: Continuous improvement	0.05	.470
1e: Lessons from accomplishments and failures	0.07	.325
2a: Leverages resources	0.08	.284
2b: School safety	–0.02	.800
2c: Complies with mandates	–0.01	.869
2d: Clear expectations for students and staff	0.01	.834
2e: Communicates effectively	0.06	.521
2f: Manages conflict	0.12*	.046
3a: School improvement initiatives	0.05	.533
3b: Aligns curricula and instruction	0.05	.412
3c: High-quality instruction	0.12	.092
3d: High expectations for students	0.05	.345
3e: Maximizes instructional time	0.01	.869
4a: Parent and community involvement	0.10	.169
4b: Professionalism	–0.03	.722
4c: Supports professional growth	–0.02	.797

* Significant at $p < .05$; ** significant at $p < .01$.

Note: Analyses are based on a value-added measure that combines all subjects. Recently hired principals are those who began their current positions in 2008/09 or later.

Source: Authors' calculations based on Framework for Leadership pilot evaluation data from 2013/14, student achievement and background data, and principals' job assignment data provided by the Pennsylvania Department of Education.

Table H17. Predicted difference in component scores on the Framework for Leadership in the 2013/14 pilot year between recently hired principals at the 84th and 50th percentiles of value-added in science in the 2012/13 pilot year

Component	Estimate	*p* value
1a: Strategic goals	−0.04	.548
1b: Data for decisionmaking	0.14**	.005
1c: Empowering work environment	0.09	.225
1d: Continuous improvement	−0.05	.403
1e: Lessons from accomplishments and failures	0.05	.403
2a: Leverages resources	−0.09	.109
2b: School safety	−0.01	.860
2c: Complies with mandates	−0.12*	.012
2d: Clear expectations for students and staff	−0.03	.508
2e: Communicates effectively	−0.03	.658
2f: Manages conflict	0.01	.926
3a: School improvement initiatives	−0.08	.118
3b: Aligns curricula and instruction	0.01	.852
3c: High-quality instruction	0.00	.970
3d: High expectations for students	0.06	.215
3e: Maximizes instructional time	−0.10	.097
4a: Parent and community involvement	0.00	.944
4b: Professionalism	−0.07	.309
4c: Supports professional growth	0.00	.948

* Significant at $p < .05$; ** significant at $p < .01$.

Note: Analyses are based on a value-added measure that combines all subjects. Recently hired principals are those who began their current positions in 2008/09 or later.

Source: Authors' calculations based on Framework for Leadership pilot evaluation data from 2013/14, student achievement and background data, and principals' job assignment data provided by the Pennsylvania Department of Education.

Table H18. Association between the Framework for Leadership scores in the 2012/13 pilot year and the value-added estimates for recently hired principals in the 2013/14 pilot year

Outcome	Value-added measure	Predicted difference in FFL score between principals at the 84th and 50th percentiles of value-added		Number of school leaders
		Estimate	*p* value	
Full Framework for Leadership score	All subjects	0.03	.531	101
	Math	0.03	.413	101
	Reading/writing	0.00	.953	101
	Science	0.03	.379	101
Score on domain 1 (Strategic/cultural leadership)	All subjects	–0.03	.525	101
	Math	–0.01	.891	101
	Reading/writing	–0.05	.225	101
	Science	0.00	.983	101
Score on domain 2 (Systems leadership)	All subjects	0.05	.347	101
	Math	0.04	.366	101
	Reading/writing	0.03	.526	101
	Science	0.04	.324	101
Score on domain 3 (Leadership for learning)	All subjects	0.07	.117	101
	Math	0.07	.121	101
	Reading/writing	0.03	.478	101
	Science	0.05	.289	101
Score on domain 4 (Professional and community leadership)	All subjects	0.01	.780	101
	Math	0.01	.753	101
	Reading/writing	–0.01	.762	101
	Science	0.05	.289	101

FFL is Framework for Leadership.

Note: Recently hired principals are those who began their current positions in 2008/09 or later. No values were statistically significant.

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

Table H19. Predicted difference in component scores on the Framework for Leadership in the 2012/13 pilot year between recently hired principals at the 84th and 50th percentiles of value-added in all subjects combined in the 2013/14 pilot year

Component	Estimate	*p* value
1a: Strategic goals	–0.02	.733
1b: Data for decisionmaking	0.08	.117
1c: Empowering work environment	0.03	.680
1d: Continuous improvement	–0.01	.859
1e: Lessons from accomplishments and failures	–0.07	.232
2a: Leverages resources	0.05	.488
2b: School safety	0.01	.937
2c: Complies with mandates	0.10*	.030
2d: Clear expectations for students and staff	0.12	.084
2e: Communicates effectively	0.04	.677
2f: Manages conflict	0.10	.133
3a: School improvement initiatives	0.09	.143
3b: Aligns curricula and instruction	0.16*	.012
3c: High-quality instruction	0.16*	.035
3d: High expectations for students	0.06	.253
3e: Maximizes instructional time	–0.01	.846
4a: Parent and community involvement	0.04	.671
4b: Professionalism	–0.03	.649
4c: Supports professional growth	0.09	.152

* Significant at $p < .05$; ** significant at $p < .01$.

Note: Analyses are based on a value-added measure that combines all subjects. Recently hired principals are those who began their current positions in 2008/09 or later.

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

Table H20. Predicted difference in component scores on the Framework for Leadership in the 2012/13 pilot year between recently hired principals at the 84th and 50th percentiles of value-added in math in the 2013/14 pilot year

Component	Estimate	*p* value
1a: Strategic goals	0.00	.970
1b: Data for decisionmaking	0.08	.114
1c: Empowering work environment	0.05	.387
1d: Continuous improvement	–0.03	.671
1e: Lessons from accomplishments and failures	0.00	.986
2a: Leverages resources	0.10	.107
2b: School safety	–0.03	.697
2c: Complies with mandates	0.08	.180
2d: Clear expectations for students and staff	0.10	.104
2e: Communicates effectively	0.10	.110
2f: Manages conflict	0.08	.208
3a: School improvement initiatives	0.08	.252
3b: Aligns curricula and instruction	0.14*	.037
3c: High-quality instruction	0.11	.127
3d: High expectations for students	0.07	.258
3e: Maximizes instructional time	0.02	.747
4a: Parent and community involvement	–0.01	.916
4b: Professionalism	–0.04	.590
4c: Supports professional growth	0.11	.058

* Significant at $p < .05$; ** significant at $p < .01$.

Note: Analyses are based on a value-added measure that combines all subjects. Recently hired principals are those who began their current positions in 2008/09 or later.

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

Table H21. Predicted difference in component scores on the Framework for Leadership in the 2012/13 pilot year between recently hired principals at the 84th and 50th percentiles of value-added in reading and writing in the 2013/14 pilot year

Component	Estimate	*p* value
1a: Strategic goals	–0.05	.378
1b: Data for decisionmaking	0.03	.546
1c: Empowering work environment	0.02	.815
1d: Continuous improvement	–0.03	.625
1e: Lessons from accomplishments and failures	–0.10*	.027
2a: Leverages resources	0.01	.884
2b: School safety	0.00	.919
2c: Complies with mandates	0.06	.193
2d: Clear expectations for students and staff	0.08	.253
2e: Communicates effectively	0.01	.896
2f: Manages conflict	0.06	.237
3a: School improvement initiatives	0.05	.380
3b: Aligns curricula and instruction	0.10	.138
3c: High-quality instruction	0.09	.196
3d: High expectations for students	0.02	.628
3e: Maximizes instructional time	–0.02	.727
4a: Parent and community involvement	0.00	.968
4b: Professionalism	–0.07	.317
4c: Supports professional growth	0.06	.249

* Significant at $p < .05$; ** significant at $p < .01$.

Note: Analyses are based on a value-added measure that combines all subjects. Recently hired principals are those who began their current positions in 2008/09 or later.

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

Table H22. Predicted difference in component scores on the Framework for Leadership in the 2012/13 pilot year between recently hired principals at the 84th and 50th percentiles of value-added in science in the 2013/14 pilot year

Component	Estimate	*p* value
1a: Strategic goals	–0.01	.874
1b: Data for decisionmaking	0.07	.134
1c: Empowering work environment	0.00	.980
1d: Continuous improvement	0.04	.524
1e: Lessons from accomplishments and failures	–0.07	.264
2a: Leverages resources	–0.01	.927
2b: School safety	0.05	.472
2c: Complies with mandates	0.09*	.043
2d: Clear expectations for students and staff	0.07	.294
2e: Communicates effectively	0.01	.923
2f: Manages conflict	0.07	.208
3a: School improvement initiatives	0.07	.286
3b: Aligns curricula and instruction	0.15**	.010
3c: High-quality instruction	0.18**	.009
3d: High expectations for students	0.03	.602
3e: Maximizes instructional time	–0.08	.270
4a: Parent and community involvement	0.13	.062
4b: Professionalism	0.07	.276
4c: Supports professional growth	0.00	.974

* Significant at $p < .05$; ** significant at $p < .01$.

Note: Analyses are based on a value-added measure that combines all subjects. Recently hired principals are those who began their current positions in 2008/09 or later.

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

Notes

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. All four domains of the Framework for Teaching in Pennsylvania had acceptable internal consistency in 2011/12, with Cronbach's alpha values ranging from 0.72 to 0.78 (Walsh & Lipscomb, 2013).
4. The average full FFL score and distribution of full scores for the subsample of school leaders for whom value-added can be estimated are identical to the average and distribution for the full sample described earlier in the report.
5. 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.
6. The errors-in-variable regression adjustment was not applied for students who repeated a grade because the samples of such students were too small. Students with very rare grade progressions—for example, students who appeared to move into a lower grade—were excluded from the VAMs.
7. Missing values of the student characteristics in X_{iy} were also imputed.
8. 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.
9. 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.
10. 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.
11. The correlations are adjusted by scaling them by the inverse of the square root of the estimated reliability of the value-added estimates. This reliability is calculated using the estimated standard errors on the value-added estimates. See Jacob and Lefgren (2008) for details on this method.

References

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

Brophy, J. E., Coulter, C. L., Crawford, W. J., Evertson, C. M., & King, C. E. (1975). Classroom observation scales: Stability across time and context and relationships with student learning gains. *Journal of Educational Psychology, 67*(6), 873–881.

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

Chiang, H., Lipscomb, S., & Gill, B. (forthcoming). Is school value-added indicative of principal quality? *Education Finance and Policy*.

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

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd edition). Hillsdale, NJ: Erlbaum.

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 edition). Crows Nest, Australia: Allen & Unwin.

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

Dhuey, E., & Smith, J. (forthcoming). How important are school principals in the production of student achievement? *Canadian Journal of Economics*.

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. C., Murphy, J., Porter, A. C., & Elliot, S. N. (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. (2015). Using student test scores to measure principal performance. *Educational Evaluation and Policy Analysis*, *37*(1), 3–28. http://eric.ed.gov/?id=EJ1050959

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

Jacob, B. A., & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics, 26*(1), 101–136.

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. 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.

Polikoff, M. (2015). The stability of observational and student survey measures of teaching effectiveness. *American Journal of Education, 121*(2), 183–212.

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, NY: 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

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. http://eric.ed.gov?id=ED550494.

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