Stated Briefly

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

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

This study analyzes the score variation, internal consistency, score stability, and concurrent validity of the Framework for Leadership (FFL)—Pennsylvania's tool for evaluating school leaders' effectiveness—during its pilot implementation in 2012/13 and 2013/14.

Why this study?

States and districts across the country are revising how they evaluate school principals, but they face a substantial challenge: there is scant evidence on the validity and reliability of current principal evaluation tools.

This Stated Briefly report is a companion to two reports:

- 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://ies.ed.gov/ncee/edlabs
- McCullough, M., Lipscomb, S., Chiang, H., Gill, B., & 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. http://ies.ed.gov/ncee/edlabs

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

Pennsylvania is among the states 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 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, known as components, as distinguished, proficient, needs improvement, or failing (see appendix A for details on the structure of the FFL). The practices are grouped into four domains: strategic/cultural leadership, systems leadership, leadership for learning, and professional and community leadership. The evaluation tool was piloted in 2012/13 and 2013/14 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 to compile statistical evidence on how well FFL scores measure school leaders' effectiveness. This study examines four descriptive research questions about key properties of the FFL:

1. *To what extent do full FFL, domains, and component 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.

2. *What is the internal consistency of the full FFL and its domains?* Internal consistency—the degree to which different parts of the FFL lead to similar conclusions about a school leader's effectiveness—is desirable because the leadership qualities captured by different parts are supposed to reflect an overall capability to improve student achievement through effective school leadership.

3. *How stable are full FFL, domains, and component scores across years?* Year-to-year stability in scores for the same school leader is important because high instability could suggest low reliability.

4. *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 FFL scores.

An interim report examined the first, second, and fourth research questions using data from the 2012/13 pilot year (Teh, Chiang, Lipscomb, & Gill, 2014). A final report examined all four research questions using data primarily from the 2013/14 pilot year (McCullough, Lipscomb, Chiang, Gill, & Cheban, 2016). This brief summarizes the findings of both reports.

What the study found

Key findings include:

- Most school leaders received scores of proficient or distinguished (the top two of four performance categories) in each practice measured by the FFL.
- 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. In other words, principals' estimated value-added scores were positively related to their FFL scores.

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

On every component, most principals and assistant principals received a rating of proficient or distinguished (see appendix A for a detailed description of components.) 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 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). The proportions of proficient and distinguished component ratings were nearly identical to those in the 2012/13 pilot year.

Scores for each domain and the full Framework for Leadership were similarly limited in variation. Consistent with the prevalence of high component score ratings, domain scores assigned to school leaders were overwhelmingly likely to equal 2.0 (corresponding to proficient) or above on the 0–3 point scale. Likewise, full FFL scores were concentrated at the top third of the rating scale in both pilot years.

The prevalence of high scores among school leaders could have occurred if highly effective leaders were most likely to participate in the pilot. However, leaders participating in the pilot made contributions to student achievement growth that varied substantially and were indistinguishable from the contributions of nonparticipating 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 full FFL had good internal consistency for principals in both pilot years (based on Cronbach's alpha values of .88 in 2012/13 and .90 in 2013/14). The internal consistency for assistant principals was good in 2012/13 and acceptable in 2013/14 (based on Cronbach's alpha values of .85 in 2012/13 and .79 in 2013/14). For both types of school leaders, the main conclusion about internal consistency from the pilot is that the different domains yield similar assessments of a school leader's effectiveness.

Framework for Leadership scores were moderately stable across years

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.

Full FFL scores for principals were moderately stable during the two pilot years (figure 1). The correlation coefficient was .54 across the full sample of principals participating in both pilot years, which is consistent with other findings on the stability of teacher observation instruments (Brophy, Coulter, Crawford, Evertson, & King, 1975; Polikoff, 2015). Although the sample of assistant principals with scores in both pilot years was relatively small (26), full FFL scores for assistant principals were highly stable across years (.80).

Year-to-year stability of the full FFL score was highest among principals who were rated on the same set of components by the same rater in both years (.68). However, not all principals were rated on the same set of components in both pilot years because principals and their supervisors jointly determined in each year the set of components for which sufficient evidence to assign a rating existed. The correlation coefficient for principals rated by the same supervisor in both years on a different set of components is .38. Among the sample of principals rated on the same set of components in both years by a different supervisor, the correlation coefficient is .42, which is similar to the observed year-to-year stability of teacher observations for the Measures of Effective Teaching study, which used multiple raters (Polikoff, 2015).

Higher Framework for Leadership scores were associated with larger estimated contributions to student achievement growth in some subjects but not in others in the 2013/14 pilot year

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 study examined correlations

Figure 1. 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.

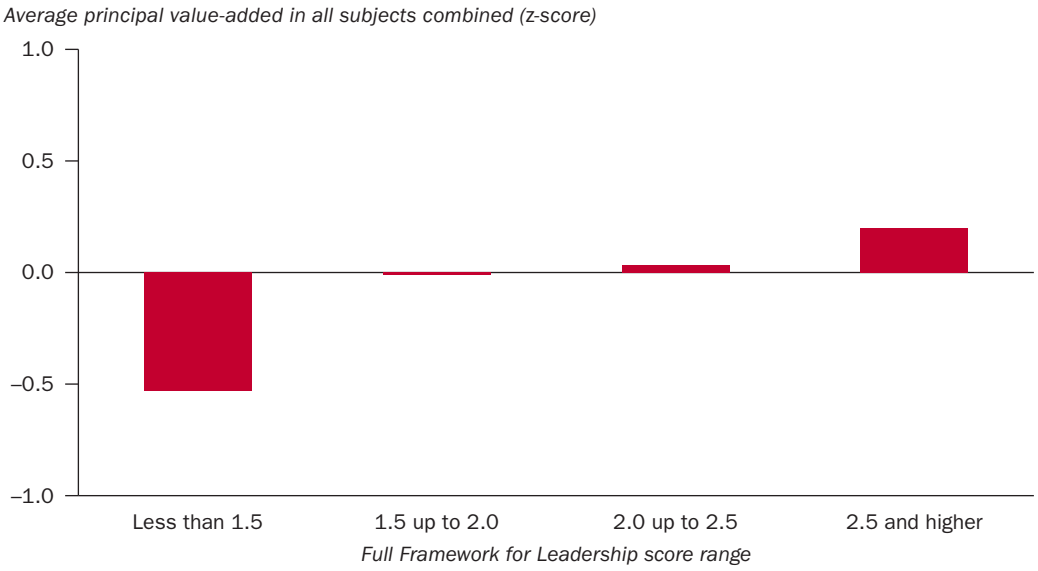
of school leaders' scores with value-added measures of their contributions to student achievement growth in the same year. The study's measure of principal value-added is based on how the school's value-added deviates from its predicted value-added, which is based on its value-added before the current 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. Because of data constraints, the relationships between value-added and FFL scores were estimated only for principals who began leading their current schools in 2008/09 or later.

Higher full FFL scores were significantly associated with higher value-added in math and marginally significantly ($p < .10$) associated with higher value-added in science and in all subjects combined in the 2013/14 pilot year. (No relationship was found between FFL scores and value-added estimates in the 2012/13 pilot year.) In other words, principals with higher estimated contributions to student achievement growth in math tended to have higher full FFL scores, and there were marginally significant tendencies for principals with higher value-added in science and in all subjects combined 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.

FFL scores most clearly differentiate among principals in terms of their value-added at the highest and lowest ranges of the scale (figure 2)—which might be where differentiation is most important.

Higher scores in FFL domain 2 (systems leadership) and domain 4 (professional and community leadership) were associated with larger value-added in all subjects combined. Scores in domains 2 and 4 were also positively related to estimated value-added in both math and science. No association with value-added in reading or writing was detected for any domain score.

Figure 2. 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 principals' job assignment data provided by the Pennsylvania Department of Education.

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. 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, which is why these correlational analyses included an examination of grade-specific subgroups. The smaller sample sizes for this analysis also made it more difficult to detect statistically significant relationships.

Implications of the study

The findings from the pilot indicate that the FFL is a promising principal evaluation tool. A key strength is its reliability, as measured by both internal consistency and year-to-year stability. While no relationship between school leaders' FFL scores and school leaders' value-added was found during the 2012/13 pilot year, the 2013/14 pilot year provided the first evidence of concurrent validity: FFL scores differentiate principals who make larger or smaller contributions to student achievement growth. This evidence of concurrent validity sets the FFL apart from other principal evaluation tools.

One area where additional examination of the FFL may be warranted, particularly during wider implementation, is the distribution of scores. 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. Scores were of low stakes during the pilot years. The variation may become even more compressed when scores become part of formal evaluations if the high stakes incentivize lenient ratings. The Pennsylvania Department of Education and other states and districts interested in implementing the FFL could consider providing ongoing training and guidance to promote consistency among supervisors assigning ratings and accurate, uniform interpretation of the rating categories.

The Pennsylvania Department of Education could also continue to gather evidence on the statistical properties of the FFL as the instrument is implemented widely. Monitoring wider implementation would 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 would enable the Pennsylvania Department of Education to examine additional measures of validity and reliability and to refine the FFL as needed.

Appendix A. 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 A1). In the pilot evaluations school leaders decided jointly with their supervisors which FFL components to use but were required to be rated on at least two components within every domain. 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. The ratings supervisors assign are based on direct observation and on evidence submitted by the school leaders.

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 evaluations assigned only component scores. For the analysis the study team computed a school leader's domain score as the equal-weighted average of scores from the components on which the leader was evaluated in that domain. The Pennsylvania Department of Education regards the four domains as equally weighted elements of a school leader's annual evaluation rating, so the study defined a school leader's full FFL score as the equal-weighted average of the four domain scores.

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 limited to the 19 other components that were consistent across the 2012/13 and 2013/14 pilot years.

Table A1. 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.

(continued)

Table A1. Components of the Framework for Leadership, by domain *(continued)*

Name of component	Description of component
2d. Establishes and implements expectations for students and staff	The school leader establishes and implements clear expectations, structures, rules, and procedures for students and staff.
2e. Communicates effectively and strategically	The school leader strategically designs and utilizes various forms of formal and informal communication with all staff and stakeholders.
2f. Manages conflict constructively	The leader effectively and efficiently manages the complexity of human interactions and relationships, including those among and between parents/guardians, students, and staff.
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.

Appendix B. Study data and methods

This appendix discusses the data and methods used in the study.

Data

The data for the study consisted of school leaders' scores on the Framework for Leadership (FFL), school leaders' job assignments and background characteristics, and student achievement scores and background characteristics.

The study used FFL scores from the end of the 2012/13 pilot year for 336 principals and 89 assistant principals and 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 evaluations assigned only component scores. For the analysis the study team computed a school leader's domain score as the equal-weighted average of scores from the components on which the leader was evaluated in that domain. The Pennsylvania Department of Education regards the four domains as equally weighted elements of a school leader's annual evaluation rating, so the study 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

Analyses to address the research question on score variation described the distributions of FFL scores for 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. The distributions of scores for domain and full 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). 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. 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 2012/13 and 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 2012/13 (for the interim report) or in 2013/14 (for the final report), 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.) The final step was to estimate a regression model for the relationship between recently hired principals' FFL scores from the end of the 2012/13 school year (interim report) or the end of the 2013/14 school year (final report) and their estimated value-added in the same year.

References

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.

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

McCullough, M., Lipscomb, S., Chiang, H., Gill, B., & 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. http://ies.ed.gov/ncee/edlabs

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

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://ies.ed.gov/ncee/edlabs

REL 2016–111

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. & Cheban, I. (2016). *Stated Briefly: Measuring school leaders' effectiveness: Findings from a multiyear pilot of Pennsylvania's Framework for Leadership* (REL 2016–111). 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.

The Regional Educational Laboratory Program produces 7 types of reports