September 2016

Can student test scores provide useful measures of school principals' performance?

Hanley Chiang

Moira McCullough

Stephen Lipscomb

Brian Gill

Mathematica Policy Research

Key findings

- Two widely feasible test-based performance measures that do not account for students' past achievement provided no information for predicting principals' contributions to student achievement in the following year.
- Two widely feasible test-based performance measures that account for students' past achievement provided, at most, a small amount of information for predicting principals' contributions to student achievement in the following year.
- Averaging test-based performance measures across multiple recent years did not improve their accuracy for predicting principals' contributions to student achievement in the following year.

ies NATIONAL CENTER FOR EDUCATION EVALUATION AND REGIONAL ASSISTANCE

Institute of Education Sciences
U.S. Department of Education

U.S. Department of Education
John B. King, Jr., *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, *Project Officer*

NCEE 2016-002

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.

September 2016

This report was prepared for the Institute of Education Sciences (IES) under Contract ED-IES-12-R-0040 by the Independent Review and Evaluation for Regional Educational Laboratories project administered by Mathematica Policy Research. 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.

Chiang, H., McCullough, M., Lipscomb, S., & Gill, B. (2016). *Can student test scores provide useful measures of school principals' performance?* (NCEE 2016–002). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

This report is available on the Institute of Education Sciences website at http://ies.ed.gov/ncee.

Summary

States and districts need ways of measuring principal performance that correctly identify effective principals. Previous federal initiatives encouraged states to develop accurate measures of principals' performance, especially measures based on student achievement growth. For example, in the four years before federal enactment of the Every Student Succeeds Act (ESSA) in 2015, more than 40 states committed to use student achievement growth in annual principal evaluations as a condition for receiving enhanced flexibility under the law's predecessor. Given that the ESSA does not specify how principals should be evaluated, states must decide whether to continue using student achievement growth to evaluate principals and, if so, how to design such performance measures.

Unfortunately, existing research offers little guidance to policymakers on which types of performance measures provide valid information about principals' contributions to student achievement. States have therefore had to develop principal performance measures without clear evidence that these measures accurately identify effective principals.

The U.S. Department of Education's Institute of Education Sciences sponsored this study to examine the accuracy of test-based measures of principal performance that could be implemented broadly. Specifically, this study assessed the *predictive validity* of these measures—the extent to which ratings from these measures accurately reflect principals' contributions to student achievement in future years. Performance measures ought to have high predictive validity to be useful for informing personnel decisions about principals. Because such decisions determine which individuals will lead schools in subsequent years, states and districts need measures that accurately identify which principals are likely to perform well in the future.

This study examined four alternative measures of principal performance. **Average achievement** used only information about students' end-of-year achievement without taking into account the students' past achievement. In contrast, **school value-added** accounted for students' own past achievement by measuring their growth—specifically, the extent to which student achievement growth at a school differed from average growth statewide for students with similar prior achievement and background characteristics. Two other measures in this study took into account the schools' prior performance to avoid rewarding or penalizing principals simply for being assigned to schools that had better or worse characteristics. **Adjusted average achievement** and **adjusted school value-added** credited principals if their schools' average achievement and value-added, respectively, exceeded predictions for the average principal, given the schools' past performance on those same measures.

To assess each measure's predictive validity, the study conducted two sets of analyses using student and principal data from 2007/08–2013/14 in the entire state of Pennsylvania. First, the study assessed the extent to which ratings from each measure are stable—that is, remain consistent over time—by examining the association between principals' ratings from earlier and later years. Stability was important to measure because only stable parts of a rating have the potential to contain information about principals' future performance; unstable parts reflect only transient aspects of their performance.

Second, the study examined the relationship between the stable part of a principal's rating and his or her contributions to student achievement in future years. To do so, the study carried out a benchmark approach to obtain the most rigorous available measure of

principals' contributions—but one that was available for only a subset of principals. For the benchmark approach, the study team calculated the change in student achievement at a school when one principal replaced another to determine how the successor's contribution differed from that of the predecessor. The study then compared the stable parts of the ratings from each of the four measures to results from the benchmark approach in a future year.

Using the results of both analyses, the study summarized each measure's predictive validity by simulating its accuracy for predicting principals' contributions to student achievement in the following year. A measure could have high predictive validity only if it was highly stable between consecutive years (from the first analysis) and its stable part was strongly related to principals' contributions to student achievement (from the second analysis). The study assessed the predictive validity of single-year ratings and ratings averaged across three years.

Key findings

The study had the following key findings:

- **The two performance measures in this study that did not account for students' past achievement—average achievement and adjusted average achievement—provided no information for predicting principals' contributions to student achievement in the following year.** Although average achievement was highly stable, those stable ratings largely reflected influences on student achievement that were outside of principals' control and therefore did not predict the principals' future contributions. Adjusted average achievement was highly unstable, reflecting purely transient aspects of performance.
- **The two performance measures in this study that accounted for students' past achievement—school value-added and adjusted school value-added—provided, at most, a small amount of information for predicting principals' contributions to student achievement in the following year.** The low predictive validity of both measures was due partly to some instability and partly to inaccuracy in the stable parts of the measures. Although school value-added ratings had a statistically significant relationship with principals' future contributions, the relationship was small in magnitude. Less than one-third of each difference in school value-added ratings across principals reflected differences in their contributions in the following year. The relationship between adjusted school value-added ratings and principals' contributions in the following year was similar in magnitude but not statistically significant.
- **Averaging performance measures across multiple recent years did not improve their accuracy for predicting principals' contributions to student achievement in the following year.** For both of the value-added measures, a principal's average rating over three years did not predict his or her future contributions more accurately than did a rating from the most recent year only. Even though ratings from prior years helped offset the unstable parts of the most recent year's rating (improving predictive validity), those prior-year ratings also captured outdated aspects of performance that principals no longer demonstrated (reducing predictive validity).

Suggestions for the design of principal evaluation systems

Based on this study and other existing research, no available measures of principal performance have yet been shown to accurately identify principals who will contribute successfully to student outcomes in future years. This study found little evidence that any widely feasible test-based measures could accurately predict principals' contributions in the following year. At the same time, no research has ever determined whether nontest measures, such as measures of principals' leadership practices, predict their future contributions.

For states and districts that still need to evaluate principals in some manner, the evaluation systems should emphasize measures that provide at least some information about principals' future contributions to student outcomes. Based on the evidence from this study, the two measures that do not account for students' past achievement—average achievement and adjusted average achievement—should not be used in a principal evaluation system. Instead, principal evaluation systems that use test-based measures should emphasize ratings based on school value-added or adjusted school value-added. Those are the only two test-based measures shown to have any degree of predictive validity.

Nevertheless, even the value-added measures will make plenty of mistakes when trying to identify principals who will contribute effectively or ineffectively to student achievement in future years. Therefore, states and districts should exercise caution when using these measures to make major decisions about principals. Given the inaccuracy of the test-based measures, state and district leaders and researchers should also make every effort to identify nontest measures that can predict principals' future contributions to student outcomes.

CONTENTS

Summary i

Key findings ii

Suggestions for the design of principal evaluation systems ii

Why this study? 1

What the study examined 3

Principal performance measures included in the study 3

Research questions 4

What the study found 8

Performance ratings from a single year were rarely stable over periods longer than one year, and those based on multiyear averages would not be any more stable 8

When focusing just on stable differences in ratings across principals, average achievement provided no information to predict principals' contributions to student achievement, while school value-added and adjusted school value-added produced only partially accurate or uncertain predictions 12

When using the original, full performance measures to predict principals' contributions to student achievement in the following year, none of the measures produced accurate predictions 17

Implications and limitations of the study findings 20

Suggestions for the design of principal evaluation systems and future research 21

Limitations of the study 24

Appendix A. Data used in the study A-1

Appendix B. Technical details of the principal performance measures that the study assessed B-1

Appendix C. Technical details of methods for examining the stability of principal performance ratings C-1

Appendix D. Technical methods for assessing the predictive validity of principal performance measures D-1

Appendix E. Study sample for assessing predictive validity E-1

Appendix F. Supplemental findings F-1

Notes Notes-1

References Ref-1

Boxes

Box 1. Principal performance measures that the study examined 3

Box 2. Data and methods 6

Figures

Figure 1. Percentage of any difference in single-year ratings across principals that was stable into subsequent years, for each measure based on math test scores 10

Figure 2. Percentage of any difference in single-year ratings across principals that was stable into subsequent years, for each measure based on reading test scores 10

Figure 3. Percentage of any difference in value-added ratings across principals that was stable to the next year, according to whether the ratings were based on a single year or a three-year average 12

Figure 4. Classification of schools according to the change in student math achievement between the departing principal's final year and the incoming principal's second year 14

Figure 5. Classification of schools according to the change in student reading achievement between the departing principal's final year and the incoming principal's second year 14

Figure 6. Relationships between the stable part of each measure and principals' future contributions to student achievement (prediction coefficients) 16

Figure 7. Relationships between the original single-year measures and principals' contributions to student achievement in the next year (prediction coefficients) 19

Figure 8. Relationships between three-year rolling averages of performance ratings and principals' contributions to student achievement in the next year (prediction coefficients) 20

Figure D1. Combinations of measurement and validation periods the study examined D-8

Figure D2. Example of a principal comparison in the study D-9

Figure F1. Percentage of principals whose single-year performance ratings were statistically distinguishable or indistinguishable from the average performance F-3

Tables

Table 1. Classification of principal performance measures according to whether they account for past student performance and past school performance 4

Table B1. Estimated relationships between school value-added and the same measure obtained two years earlier B-11

Table B2. Years in which adjusted school value-added was estimated B-13

Table B3. Estimated relationships between average achievement and the same measure obtained two years earlier B-15

Table B4. Estimated relationships between principal performance measures and the same measures obtained in different years B-17

Table D1. Expected margin of error in estimates of prediction coefficients D-11

Table E1. How the study constructed the principal sample for the predictive validity analysis E-2

Table E2. Characteristics of Pennsylvania principals, statewide during 2010/11, and for the predictive validity analysis sample during the year in which they entered or departed E-3

Table E3. Characteristics of Pennsylvania students during 2010/11, statewide and in schools where the study estimated principals' contributions to student achievement E-4

Table F1. Descriptive statistics on the principal performance ratings F-1

Table F2. Descriptive statistics on the stable parts of the principal performance ratings F-2

Table F3. Correlations among performance measures in math for principals in the predictive validity analysis sample F-4

Table F4. Correlations among performance measures in reading for principals in the

predictive validity analysis sample F-4
Table F5. Correlations among performance measures in math for all principals in their second year or later at their school F-5
Table F6. Correlations among performance measures in reading for all principals in their second year or later at their school F-5
Table F7. Prediction coefficient estimates from the main findings F-6
Table F8. Prediction coefficient estimates for stable parts of the performance measures: supplementary findings F-6
Table F9. Prediction coefficient estimates for the original performance measures: supplementary findings F-7
Table F10. Prediction coefficient estimates from an alternative model specification that accounts for varying dips in school performance during principal transitions F-8
Table F11. Extent to which the stable part of principal performance measures predicted students' prior achievement F-11
Table F12. Extent to which the stable parts of principal performance measures predicted students' background characteristics F-12
Table F13. Extent to which the stable parts of the principal performance measures predicted students' prior achievement in the restricted sample F-13
Table F14. Extent to which the stable parts of the principal performance measures predicted students' background characteristics in the restricted sample F-14
Table F15. Prediction coefficient estimates from the restricted sample F-15

Why this study?

States and districts need ways of measuring principal performance that correctly identify effective principals. Effective principals can serve their schools in various ways, one of which is to promote better student outcomes, an aim they might achieve by attracting and retaining effective teachers, developing teachers' instructional skills, and setting clear expectations for the performance of students and staff (Grissom, Loeb, & Master, 2013; Loeb, Kalogrides, & Béteille, 2012; Purkey & Smith, 1983). Several studies have documented that students score higher on state assessments when their schools are led by more effective principals (Branch, Hanushek, & Rivkin, 2012; Chiang, Lipscomb, & Gill, 2016; Coelli & Green, 2012; Dhuey & Smith, 2013, 2014; Grissom, Kalogrides, & Loeb, 2015).

Previous federal initiatives encouraged states to develop and implement valid measures of principals' performance, especially measures based on student achievement growth. In the four years before federal enactment of the Every Student Succeeds Act (ESSA)[1] in 2015, more than 40 states committed to use student achievement growth in annual principal evaluations as a condition for receiving enhanced flexibility under the law's predecessor. Given that the ESSA does not specify how principals should be evaluated, states must decide whether to continue using student achievement growth to evaluate principals and, if so, how to design such performance measures.

One obvious way of factoring student achievement growth into principals' evaluations is to use performance measures based on student test scores, which are the focus of this report. Academic tests represent only one of several approaches to measuring student achievement, and these tests capture only particular dimensions of achievement. Nevertheless, test scores provide one of the most widely available measures of student achievement and hence were the basis for the performance measures in this study. To simplify the discussion, this report uses "student test scores" and "student achievement" synonymously.

When evaluating principals based on student achievement data, the objective is to distinguish principals who make larger and smaller contributions to student achievement (recognizing that this is not the only aim of effective school leaders). Identifying principals who are effective at raising student achievement, however, is more difficult than identifying teachers who are effective at raising student achievement. For principals and teachers alike, measuring contributions to student achievement requires accounting for differences in the students served by different educators. In the case of principals, however, there is an additional challenge because different principals lead different schools that might be dissimilar on any number of factors outside of their control. Not only might different schools serve different kinds of students, but they could also be staffed by different kinds of teachers for reasons beyond the current principals' control. For example, a principal's predecessor might have made hiring decisions that cannot be easily reversed, or a school might consistently have trouble attracting qualified teachers due to being in an unsafe

[1] The ESSA was enacted as Public Law Number 114-95 on December 10, 2015.

neighborhood. Schools could also differ in resource allocations set by district authorities. In consequence, the principal's contribution to student achievement could be quite different from the school's contribution to student achievement.

These challenges suggest that a principal performance measure must carefully distinguish a principal's contribution from the influence of other school-level factors to be an accurate evaluation tool. Conceptually, certain measures might be expected to fall short of that objective. For example, average end-of-year achievement at a school could more strongly reflect students' socioeconomic backgrounds rather than educators' contributions. Focusing on the *growth* that students make could better account for the fact that some students enter a school with better advantages than others, but average achievement growth at a school might still reflect the influence of *all* staff, resources, and amenities of a school—not just the principal. In fact, a prior study found that schoolwide student achievement growth was poorly indicative of principals' longer-run effectiveness at raising student achievement (Chiang et al., 2016). More sophisticated measures that account for schools' prior circumstances may, in theory, more accurately capture principals' contributions.

Unfortunately, existing research offers little guidance to policymakers on which measures provide valid information about principals' true contributions to student achievement. Most research on principals' contributions to student achievement has focused on demonstrating that student achievement at a school typically changes when its principal is replaced by another (Branch et al., 2012; Coelli & Green, 2012; Dhuey & Smith, 2013, 2014). Although this research suggests that principals differ in effectiveness from their predecessors or successors, it offers few lessons for how a state could design evaluation measures that could be used for all principals. Other studies have documented the extent to which different types of principal performance measures are related to each other (Grissom et al. 2015; McCullough, Lipscomb, Chiang, Gill, & Cheban, 2016; Teh, Chiang, Lipscomb, & Gill, 2014), but those relationships do not reveal which measures most faithfully reflect principals' contributions to student achievement.

With few insights from existing research, states have had to proceed with developing principal performance measures based on student achievement data without clear evidence that these measures are valid indicators of principals' contributions to student achievement. As of 2013, most states that used student achievement data to evaluate principals were relying on measures of schoolwide student achievement growth that made no attempt to distinguish principals' contributions from school effects that are outside of principals' control (Goldring & Jones, 2014). Yet, to date, there is no evidence that any alternative measures would be more valid.

To fill this void of evidence, this study, commissioned by the U.S. Department of Education's Institute of Education Sciences, examined the validity of test-based principal performance measures. Specifically, this study used student and principal data from the entire state of Pennsylvania to assess the *predictive validity* of test-based measures—the extent to which ratings from these measures accurately reflect principals' contributions to student achievement in future years. Performance measures ought to have high predictive validity to be useful for informing personnel decisions about principals. Because such decisions determine which individuals will lead schools in subsequent years, states and districts need measures that accurately identify which principals are likely to perform well in the future.

This study is the first to assess the predictive validity of several test-based measures of principal performance. Lessons from this study can help states choose and improve their approaches for using student achievement data to evaluate principals.

What the study examined

This study assessed the validity of different performance measures that states could feasibly and widely implement to evaluate principals based on student test score data. All measures had a common objective: to accurately identify principals who make larger and smaller contributions to student achievement. The study's central aims were to determine the extent to which each measure fulfilled this objective and examine the characteristics of the measures that influenced how well they fulfilled this objective.

Principal performance measures included in the study

The study examined four types of principal performance measures: average achievement, adjusted average achievement, school value-added, and adjusted school value-added (see box 1 for a description of each measure). The study selected these measures because they fulfilled two main requirements that made them potentially suitable for implementation in a large-scale evaluation system. First, states could use these measures to evaluate most principals who lead schools with tested grades and subjects. Second, these measures could compare any principal with any other in the principal's state.

Box 1. Principal performance measures that the study examined

Average achievement: The average end-of-year achievement among students at the principal's school.

Adjusted average achievement: The extent to which average achievement at a school differs from what would be predicted for the average principal, given the past average achievement of earlier student cohorts at the school.

School value-added: The extent to which student achievement growth at a school differs from average growth statewide for students with similar prior achievement and background characteristics.

Adjusted school value-added: The extent to which a school's value-added differs from what would be predicted for the average principal, given the past value-added for earlier student cohorts at the school.

Aside from those two basic similarities, the principal performance measures this study examined represent very different ways of trying to measure principals' contributions to student achievement (table 1). Average achievement uses only information about students' end-of-year test scores without taking into account students' past test-score history. The academic proficiency rate, a measure widely used for school accountability, is an average achievement measure similar in spirit to the one examined in this study. In contrast, school value-added accounts for students' own past achievement by measuring their *growth*—specifically, the extent to which student achievement growth at a school differs from average growth statewide for students with similar prior achievement and background characteristics.

The measures in this study also differ in whether they account for schools' prior performance with earlier cohorts of students. Some schools might generate persistently better achievement or achievement growth than other schools due to factors beyond principals' control. For example, some schools might consistently attract better teachers because they are closer to neighborhoods where those teachers live. Two measures in this study take into account the schools' prior performance as a way to avoid rewarding or penalizing principals simply for being assigned to schools that have better or worse characteristics. In particular, adjusted average achievement credits principals if their school's current rating for average achievement exceeds predictions for the average principal, given the school's past rating for average achievement. Likewise, adjusted school value-added credits principals if their school's current value-added exceeds predictions for the average principal, given the school's past value-added.

Table 1. Classification of principal performance measures according to whether they account for past student performance and past school performance

	Does not account for prior school performance	Assesses current school performance relative to predictions based on prior school performance
Does not account for students' prior test scores	Average achievement	Adjusted average achievement
Uses students' prior test scores to measure student growth	School value-added	Adjusted school value-added

Source: Authors' compilation.

On each of the four measures, this study used data on the reading and math scores of students in Pennsylvania to generate single-year performance ratings for principals in the state from 2009/10 to 2013/14 (see box 2 and appendix A for a description of the data, and appendix B for technical details on how the study constructed the performance ratings). The study refers to these ratings as single-year ratings because they were based on final student test score outcomes from a single year. As discussed earlier, several measures compared those final outcomes to predictions based on earlier achievement data, but the outcomes themselves still came from one year. These performance ratings contributed only to this study's analyses and were not used for actual evaluations in Pennsylvania.

Research questions

Many of the ways in which states or districts could use performance ratings in personnel decisions assume that the ratings provide accurate information about principals' future contributions to student achievement. For example, decisions to retain principals with high performance ratings assume that these principals will contribute effectively to student achievement in subsequent years. Likewise, policies that provide extra assistance or professional development for principals with low performance ratings assume that they would continue to perform poorly without those supports. Therefore, useful performance measures for states or districts should provide accurate information for assessing principals' future contributions to student achievement—that is, they should have high predictive validity.

The study examined the predictive validity of the four performance measures by addressing the following two questions (see box 2 for an overview of the analytic approach).

Question 1: On each performance measure, how stable—that is, consistent over time—are differences in performance ratings across principals? Determining a performance measure's predictive validity requires first knowing how much of the variation in ratings is stable. Only stable differences in ratings between principals have any potential for predicting principals' future contributions to student achievement. Unstable differences—those that disappear in subsequent years—capture transient aspects of principals' performance that have no bearing on their contributions in the future.[2] A valid measure need not be perfectly stable, however, given that principals' performance might truly evolve from learning on the job.

Question 2: How accurately do the performance ratings from each measure predict principals' contributions to student achievement in future years?[3] Stability alone does not guarantee that a performance measure has predictive validity. Ultimately, performance measures must capture the information they are intended to measure: principals' future contributions to student achievement.

The predictive validity of a performance measure depends on both how much of the measure is stable and how closely that stable part reflects principals' contributions in subsequent years. Therefore, given the findings on stability from the first research question, the study assessed the predictive validity of each test-based performance measure in two stages. First, the study examined the accuracy of the stable part of each measure—specifically, the extent to which stable differences in ratings were associated with principals' future contributions to student achievement.

Second, the study combined findings on both the stability of each measure and the accuracy of its stable part to summarize the predictive validity of the measure as a whole. Specifically, the study simulated how accurately each original, full measure could predict principals'

[2] This study examined the stability of principals' performance ratings regardless of whether they remained at the same schools across years. All measures are intended to ascertain the effectiveness of principals—not schools—so the ratings are supposed to gauge how well principals would perform at whatever schools they are assigned.

[3] For a performance measure to make good predictions about principals' future contributions, it must accurately reflect the part of a principal's contribution that remains consistent over time. Therefore, measures that accurately predict principals' future contributions must also do a good job of predicting principals' past contributions. Accordingly, the study's analytic approach examined how well performance ratings predict principals' contributions to student achievement in another (past or future) year. However, because making predictions about future performance is the goal in practice, this report uses the term "future year," for simplicity, to describe a year outside of the period during which the performance rating was determined, whether a prior year or a subsequent year.

contributions in the following year. This summary analysis focused on each full measure—not just its stable part—because, in practice, a state or district may not want to focus only on stable ratings, which do not capture ways in which a principal's performance truly evolves over time. Moreover, this analysis focused on predicting principals' contributions in the following year because this would be the most relevant information for end-of-year decisions on which principals to retain, assist, or replace in the coming year.

When addressing these research questions, the study examined two versions of each measure: ratings from a single year and ratings averaged over three years. Examining both single-year and three-year ratings provided insight on whether averaging across multiple years could improve the measures' stability and predictive validity.

Box 2. Data and methods

Data

The study used administrative data on all students and principals across the state of Pennsylvania (see appendix A for details). The Pennsylvania Department of Education supplied all of the data for this study.

Data on students' test scores and background characteristics were the key inputs into ratings of principals' performance. The test score data consisted of reading and math scores on the Pennsylvania System of School Assessment for all of Pennsylvania's students in grades 3 to 8 in each year from 2006/07 to 2013/14. The student background data, which enabled some principal performance measures to control for student demographic characteristics, were available from 2007/08 to 2013/14.

Data on principals' job assignments enabled the study to link each school's student achievement data to the principal who led the school. These data identified the principal who led each school in Pennsylvania in each year from 2007/08 to 2013/14.

The analysis sample was restricted to principals who were in their second year or later as their school's leader, including 7,352 principal-year combinations and 2,424 distinct principals. The study excluded first-year principals because prior research suggested that principals could have only a very limited influence on student achievement in their first year (Coelli & Green, 2012). Thus, for a given principal, the principal-year combination for the principal's first year at a school was dropped from the dataset, but all subsequent principal-year combinations for the same principal at the same school were retained.

Methods

Before addressing the research questions, the key first step was to construct ratings of principals' performance based on each of the four measures this study examined (see appendix B for details). On each measure, the study generated performance ratings for principals who led schools with any grades from 5 to 8, separately in each year from 2009/10 to 2013/14. Grade 5 was the earliest grade in which most students had test scores from two years earlier, a key student characteristic that the value-added measures took into account. The 2009/10 school year was the earliest year in which data were available to measure a school's value-added from two years earlier, a key input into the adjusted school value-added measure.

(*continued*)

Box 2. Data and methods (continued)

Method for addressing research question 1 on whether the performance measures produced stable differences in ratings across principals: The study identified principals who had pairs of single-year performance ratings that were one, two, three, and four years apart (yielding sample sizes of 5,272; 3,521; 2,195; and 1,180 principal-year combinations, respectively). In each of those time intervals, the study estimated a regression model for the relationship between performance ratings from the initial and final year, with the estimated coefficient on the initial year indicating the proportion of variation in principals' ratings that was stable (constant over time) during that time interval. This analysis also produced a modified version of each performance measure, used in addressing research question 2, that isolated just the stable part of the measure by filtering out unstable differences in ratings (Chetty, Friedman, & Rockoff, 2014; see appendix B for details).

The study used findings on the stability of the single-year ratings to simulate the stability of a three-year rolling average—specifically, the proportion of the variation in the three-year average that would persist to the following year. Appendix C provides detailed formulas.

Method for addressing research question 2 on the predictive validity of each performance measure: The study identified pairs of principals in which one principal replaced the other as a school's leader. For each pair, the study defined a transition period that included the departing principal's final year through the incoming principal's second year. Sixty-seven pairs of these principals had performance ratings from a common year outside of their transition period, but the amount of time elapsing between the initial ratings and the transition period varied across pairs. The analyses for research question 2 focused on this set of 67 principal pairs.

The goal of the analyses was to assess the degree to which principals' performance ratings from outside the transition period reflected their contributions to student achievement measured during the transition period. Specifically, the change in the school's average student achievement during the transition period served as the most rigorous available benchmark measure of the difference in the principals' contributions to student achievement. That is, it provided an estimate of how achievement at the school under the new principal's leadership differed from the achievement that would be expected had the outgoing principal continued for an additional two years.

The assessment of predictive validity proceeded in two stages, as follows:

- **Examining the predictive validity of the stable part of each measure:** This stage used a regression model. For each pair of principals, the dependent variable of the model was the difference between the two principals in their contributions to student achievement (calculated from the benchmark approach during the transition period). The key independent variable was the stable difference in performance ratings between the two principals (calculated from one of the four measures outside the transition period). The estimated regression coefficient, called the prediction coefficient, captured the extent to which the stable part of the performance measure accurately predicted principals' future contributions to student achievement. The prediction coefficient was then compared to an ideal value of 1, which would indicate that the performance measure accurately predicted principals' future contributions, on average.
- **Summarizing the predictive validity of each original, full measure:** This stage combined findings from both the stability of each measure and the predictive validity of its stable part to simulate the extent to which each measure as a whole could predict principals' contributions in the following year. This overall prediction coefficient was equal to the proportion of the variation that was stable between consecutive years (from research question 1) multiplied by the prediction coefficient of the stable part of the measure (from the first stage of research question 2).

The study's precision for estimating the prediction coefficient of each original measure was similar to the precision of one prominent prior study on the predictive validity of teacher value-added measures (Kane et al., 2013) but worse than that of another study (Chetty et al., 2014). Details on the methods used for constructing benchmarks, analyzing the predictive validity of the four performance measures, and gauging the precision of those analyses are available in appendix D; details on how the study constructed the predictive validity analysis sample are available in appendix E.

What the study found

This section describes the study's findings on the stability of ratings from each principal performance measure and how accurately, on average, the stable parts of the performance measures predicted principals' future contributions to student achievement. This section then combines findings from both analyses to summarize the predictive validity of each original, full measure in a way relevant to personnel decisions: each measure's accuracy for predicting principals' contributions in the following year. As throughout the report, this section continues to refer to student achievement as shorthand for student test scores.

Performance ratings from a single year were rarely stable over periods longer than one year, and those based on multiyear averages would not be any more stable

For a performance measure to provide any information about principals' future contributions to student achievement, differences in ratings across principals must have some degree of stability—that is, some consistency over time. If the differences are highly unstable—such that principals with initially higher ratings than others would not continue to have higher ratings in subsequent years—then the measure would reflect mostly temporary aspects of principals' performance. Temporary differences in performance ratings—which could reflect random fluctuations in principals' true performance or temporary factors outside of the principals' control that influence student achievement schoolwide—cannot identify which principals will lead schools more effectively than others in future years.

Before assessing whether differences in ratings were stable, it was important to confirm that ratings differed at all across principals. The study confirmed that the measures generated real variation in performance ratings: each measure identified many principals whose performance ratings were above or below the average by a statistically significant margin—that is, a margin that should not have been due to chance alone (see appendix F for additional details). Across these measures, 20 to 41 percent of principals were statistically distinguishable as above-average performers, and 19 to 33 percent of principals were statistically distinguishable as below-average performers (see figure F1 of appendix F). The ability of each measure to reliably distinguish principals from average was consistent with that of teacher value-added measures examined in the state of Pennsylvania (Lipscomb, Chiang, & Gill, 2012).

Differences in performance ratings were moderately or highly stable for one year, but on all measures except average achievement, no more than a small fraction of each difference in ratings was stable over longer periods of time. On three single-year measures—average achievement, school value-added, and adjusted school value-added—approximately half or more of each difference in ratings persisted for one year (figures 1 and 2). In fact, nearly all (98 percent) of any difference in average achievement ratings was stable for one year. The fraction of the variation in school value-added ratings that was stable for one year (66 percent) was higher than what previous research found for teacher value-added (50 percent) (Chetty et al., 2014),[4] and the fraction

[4] In the original analyses by Chetty et al. (2014), instability could stem from both changes in teachers' true performance and fluctuations attributable to the limited numbers of students who contribute to the value-added ratings. In this study, instability

of the variation in adjusted school value-added ratings that was stable for one year (46 percent) was similar to that reference point. On the remaining measure, adjusted average achievement, approximately 35 percent of any difference was stable for one year.

However, on most measures, less than half of any difference in single-year ratings was stable for more than one year, and only a small fraction of the difference remained four years later. For example, only 15 to 25 percent of the variation in performance was stable for four years when school value-added or adjusted school value-added measures were used to measure annual performance (figures 1 and 2). In comparison, prior literature suggests that approximately 30 percent of the variation in teacher value-added ratings is stable for four years (Chetty et al., 2014). There was no stability in adjusted average achievement ratings for more than one year. In other words, if average achievement was higher or lower compared to its predicted level based on earlier cohorts, this provided no information for forecasting whether, under the same principal, average achievement would continue to be higher or lower compared to predictions two or more years later. The exception to the pattern of low stability was the unadjusted average achievement measure, which produced highly stable rating differences: 95 to 97 percent of the annual variation persisted for four years. (Details on the stability of principal rating differences across all possible time periods available in the study data are available in table B4 in appendix B.)

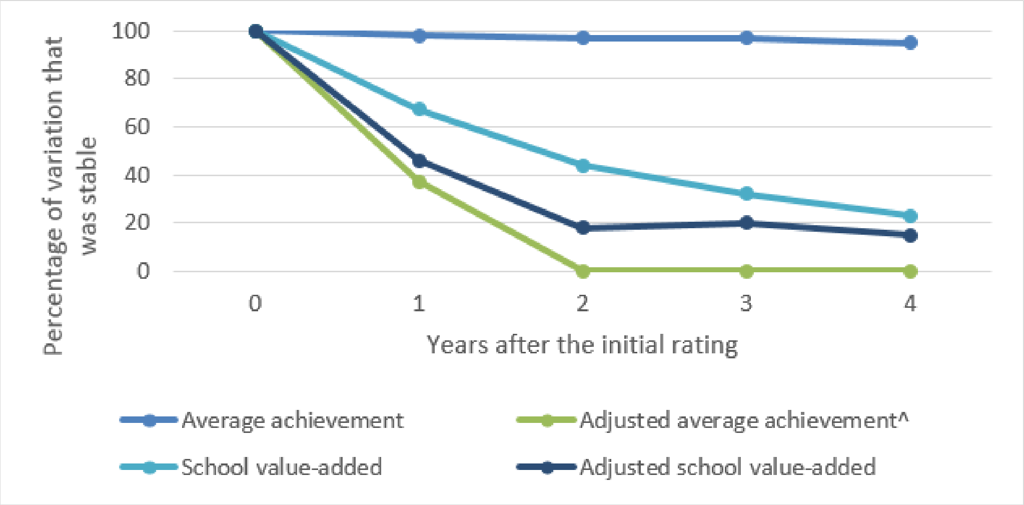
On most measures, less than half of any difference in single-year ratings was stable for more than one year, and only a small fraction of the difference remained four years later.

Principals' performance ratings might change over time for several reasons. Their actual contributions to student achievement might evolve as they learn on the job or face new challenges. Performance ratings could also change for reasons unrelated to principals' actual contributions. For example, in measures that rely on comparing achievement or achievement growth to predictions, performance ratings could fluctuate if those predictions are unduly sensitive to school characteristics that are unstable across years.

Because most single-year measures of principal performance have such limited stability over multiple years, they are unlikely to provide accurate forecasts of principals' future performance over a long period of time. Nevertheless, the moderate stability of most measures for one year suggests that the measures have the potential to predict principals' contributions to student achievement in the following year. Findings reported later in this section will assess whether this potential was met.

from limited numbers of students has already been removed. Therefore, this study applied a similar adjustment to the values provided by Chetty et al. (2014) before comparing them with this study's findings. See appendix C for details.

Figure 1. Percentage of any difference in single-year ratings across principals that was stable into subsequent years, for each measure based on math test scores

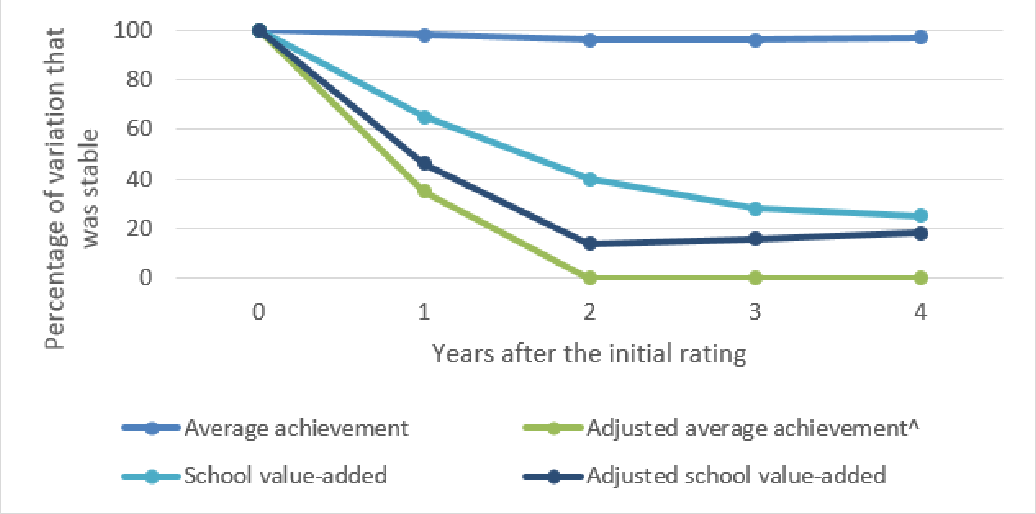


Note: Figure is based on all principals in Pennsylvania in their second year or later at their school who had performance ratings in both an initial and later year between 2008/09 and 2013/14. The numbers of principals who had ratings one, two, three, and four years apart on at least one measure were, respectively, 5,272; 3,521; 2,195; and 1,180. The sample excludes all principals from the final predictive validity analysis sample.

^ As shown in the figure, none of the differences in adjusted average achievement were stable for two or more years. In fact, those differences became reversed in direction (see appendix B, table B4).

Source: Authors' calculations based on data provided by the Pennsylvania Department of Education, 2006/7—2013/14.

Figure 2. Percentage of any difference in single-year ratings across principals that was stable into subsequent years, for each measure based on reading test scores



Note: Figure is based on all principals in Pennsylvania in their second year or later at their school who had performance ratings in both an initial and later year between 2008/09 and 2013/14. The numbers of principals who had ratings one, two, three, and four years apart on at least one measure were, respectively, 5,272; 3,521; 2,195; and 1,180. The sample excludes all principals from the final predictive validity analysis sample.

^ As shown in the figure, none of the differences in adjusted average achievement were stable for two or more years. In fact, those differences became reversed in direction (see appendix B, table B4).

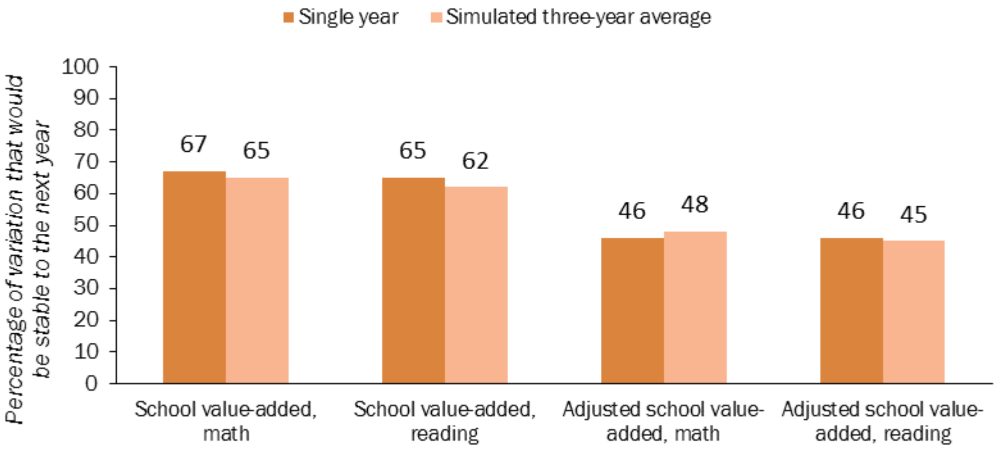
Source: Authors' calculations based on data provided by the Pennsylvania Department of Education, 2006/7—2013/14.

Using a three-year rolling average of performance ratings would not increase the stability of the value-added measures. To improve forecasts of principals' future performance on each measure, a key task is to increase the proportion of variation that is stable and decrease the proportion that is transitory or unstable. In teacher evaluations, instability of annual value-added scores has been well documented (Goldhaber & Hansen, 2008; McCaffrey, Sass, Lockwood, & Mihaly, 2009). Teacher evaluation systems sometimes incorporate a three-year rolling average of value-added scores under the assumption that doing so would enhance stability (North Carolina Department of Public Instruction, 2013; SAS Institute, Inc., 2014). The study team simulated the use of a three-year rolling average of principal performance ratings based on the unadjusted and adjusted school value-added measures to examine whether such an approach might increase the proportion of variation that is stable from one year to the next.

On each measure, the study compared the stability of a three-year rolling average of ratings with the stability of a single-year rating in the following manner. First, the study simulated the proportion of the variation in the three-year average that would persist to the following year. Given the results in figures 1 and 2, the stability of the three-year average could be simulated based on the extent to which ratings from the current year, prior year, and two years prior (the years forming the three-year average) persisted to the following year. In other words, the simulations used findings on the stability of ratings for one, two, and three years (see appendix C for technical details on the simulations). Second, the study compared the stability of the three-year average to the stability of just the current year of ratings—that is, the proportion of the variation in ratings from the current year that persisted to the following year. These comparisons did not include average achievement (because single-year performance ratings from this measure were already very stable) and adjusted average achievement (because single-year ratings from this measure were completely unstable, so combining multiple years of this measure would not make a difference).

For the value-added measures, using a three-year rolling average of performance ratings rather than just the current year would not change the percentage of the variation across principals that is stable to the next year (figure 3). For ratings measured before the current year, at least two years will have elapsed by the time the following year is completed. Because the stability of ratings declines substantially when more than one year has elapsed (see figures 1 and 2), incorporating prior ratings would not produce a measure with greater stability to the following year. In other words, even though ratings from prior years could help offset the unstable parts of the current year's rating (improving stability), those prior-year ratings would also capture outdated aspects of performance that principals would no longer demonstrate (reducing stability). This simulation suggests that, relative to using a performance rating from a single year, incorporating multiple years of performance ratings from the value-added measures would not provide any additional information to help states and districts forecast which principals will succeed or struggle on these measures in future years.

Figure 3. Percentage of any difference in value-added ratings across principals that was stable to the next year, according to whether the ratings were based on a single year or a three-year average



Note: Findings on the stability of single-year ratings are based on all principals in Pennsylvania in their second year or later at their school who had performance ratings in any two consecutive years between 2008/9 and 2013/14 (N = 5,272 principals). The sample excludes principals included in the final predictive validity analysis sample. Simulated values for the stability of a three-year average are calculated from applying the one-year, two-year, and three-year stability findings from figures 1 and 2 to the formulas shown in appendix C.

Source: Authors' calculations based on data provided by the Pennsylvania Department of Education, 2006/7—2013/14.

When focusing just on stable differences in ratings across principals, average achievement provided no information to predict principals' contributions to student achievement, while school value-added and adjusted school value-added produced only partially accurate or uncertain predictions

A performance measure is useful to states and districts only if it captures the information that it purports to measure: principals' future contributions to student achievement. In other words, policymakers need measures with high predictive validity—those that are accurate indicators of principals' contributions to student achievement in future years. This study characterized a measure as accurate if it correctly predicted principals' contributions to student achievement, on average. (In technical terms, an accurate measure is one with no forecast bias; see appendix D for technical details on forecast bias.)

For a performance measure to be accurate, a large part of the measure must be stable, and that stable part must be strongly associated with principals' future contributions to student achievement. The first condition—the stability of each measure—was discussed earlier. The discussion that follows focuses on the second condition—the relationship between the stable part of each measure and principals' future contributions.

Because unstable parts of a performance measure cannot predict principals' future contributions, the study used a statistical technique known as drift correction (Chetty et al., 2014) to identify just the stable differences in performance ratings across principals. The study then calculated the predictive validity of these stable differences (see appendix B for more detail on the drift correction).

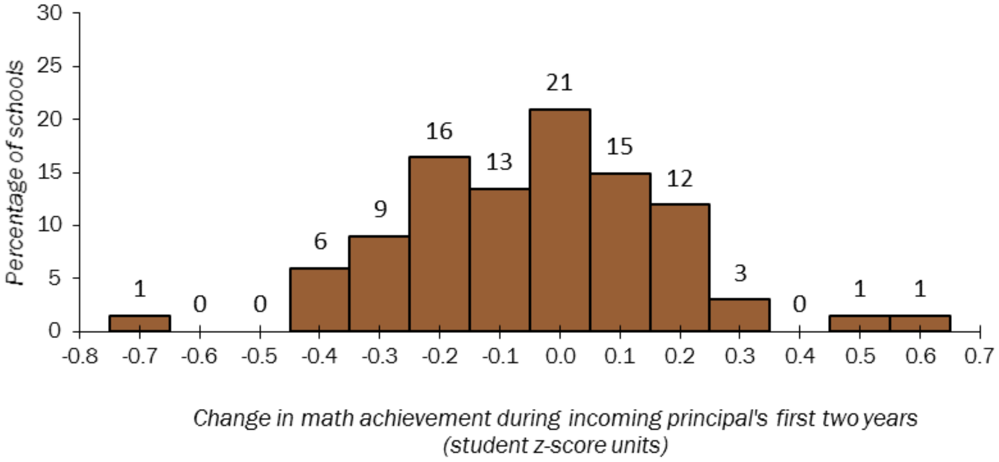
To assess the predictive validity of the stable ratings from the four performance measures, the study needed a benchmark approach—separate from any of the four measures—for measuring principals' contributions to student achievement. As noted in box 2, the benchmark approach represented the most rigorous available method for measuring principals' contributions, and the other measures were judged by how well they predicted the benchmark results in a future year. For the benchmark approach, the study team identified pairs of principals in which one principal succeeded another at the same school. For each pair, the change in the school's average student achievement during the transition period—from the predecessor's final year to the successor's second year—served as the benchmark measure of how the two principals' contributions to student achievement differed. (A detailed explanation of this approach is available in appendix D.)

This study relied on the accuracy of the benchmark approach. Because the benchmark approach compared principals who served at the same school, it held constant many factors outside of principals' control that often differ between schools—for example, neighborhood characteristics or the quality of school facilities. As a result, the benchmark approach could attribute changes in test scores during a principal transition to differences between the predecessor's and successor's contributions to student achievement. However, if other factors at a school changed in a way that was systematically related to whether a principal was replaced by a better or worse successor, then test-score changes might reflect these other factors—not just the principals' relative effectiveness. Although this study found few threats to the validity of the benchmark approach (see appendix F for evidence), there remains the potential for inaccuracy in the benchmark approach due to unmeasured changes within schools that underwent principal transitions.

At many schools with leadership transitions, student achievement changed between the departing principal's final year and the incoming principal's second year, suggesting that principals differed in their contributions to student achievement. For the predictive validity analysis to be appropriate, principals must differ in their contributions to student achievement during the transition period. The study confirmed that this was true. In both math (figure 4) and reading (figure 5), the direction and magnitude of the change in student achievement during a leadership transition varied across schools. Changes in math achievement ranged from a 0.7 standard deviation decline over a two-year period to a 0.6 standard deviation increase; the range of changes in reading achievement was identical.[5] Therefore, incoming principals typically differed from their predecessors in their contributions to student achievement.

[5] A 0.7 standard deviation decline is equivalent to moving the median student to the 24th percentile. A 0.6 standard deviation increase is equivalent to moving the median student to the 73rd percentile.

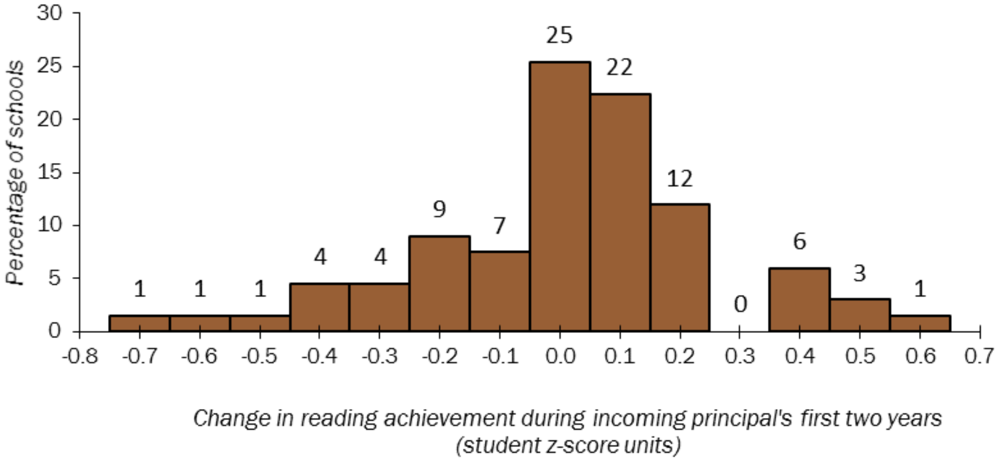
Figure 4. Classification of schools according to the change in student math achievement between the departing principal's final year and the incoming principal's second year



Note: Figure is based on 67 pairs of departing and incoming principals. For each pair, the change in student achievement was measured from the departing principal's last year to the incoming principal's second year. Across pairs, the mean of the change in achievement was –0.04 student z-score units; the standard deviation of the change in achievement was 0.22 student z-score units.

Source: Authors' calculations based on data provided by the Pennsylvania Department of Education, 2006/7—2013/14.

Figure 5. Classification of schools according to the change in student reading achievement between the departing principal's final year and the incoming principal's second year



Note: Figure is based on 67 pairs of departing and incoming principals. For each pair, the change in student achievement was measured from the departing principal's last year to the incoming principal's second year. Across pairs, the mean of the change in achievement was 0.00 student z-score units; the standard deviation of the change in achievement was 0.24 student z-score units.

Source: Authors' calculations based on data provided by the Pennsylvania Department of Education, 2006/7—2013/14.

Given that principals differed in their contributions to student achievement, how well could the stable part of each performance measure predict those differences? Examining the predictive validity of all four performance measures would make sense only if they produced at least somewhat different ratings of principal performance. The study confirmed that the measures did not always agree in their assessments of principals' performance, with a correlation of less than 0.7 for most pairs of measures (see appendix F for additional details on the correlations among measures). Overall, given that the four performance measures were not completely consistent in their assessments of principal performance, the predictive validity of the measures could differ as well.

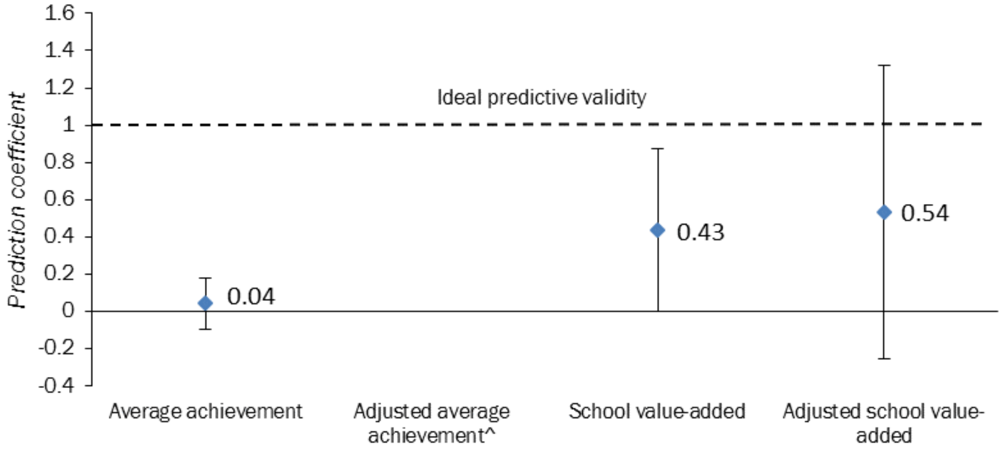
The study expressed the predictive validity of each measure with the prediction coefficient, which answered the following question: what fraction of each stable difference in performance ratings between two principals reflected the difference in their future contributions to student achievement? The standard for ideal accuracy was a prediction coefficient of 1—the scenario in which stable differences in ratings would fully reflect, on average, differences in principals' future contributions. A previous study found that stable differences in teacher value-added ratings accurately predicted teachers' future contributions to student achievement (Chetty et al., 2014). That study found a prediction coefficient of 0.974, not statistically distinguishable from 1.

Stable differences in average achievement ratings across principals provided no information about principals' contributions to student achievement in future years. The stable part of the average achievement measure had a prediction coefficient close to zero (0.04), and this coefficient was less than 1 by a statistically significant margin, confirming that the measure provided inaccurate information about principals' future contributions (figure 6). In other words, average achievement was completely uninformative about a principal's contribution to student achievement. Instead, it reflected influences on student achievement that were outside of principals' control.

Adjusted average achievement, as previously noted, had no stable part that the study could isolate to predict principals' contributions to student achievement.[6] Because this entire measure reflected purely transient aspects of performance, no part of this measure had any predictive validity.

[6] Although about 35 percent of the variation in adjusted average achievement ratings was stable for one year (see figures 1 and 2), the study's analysis of predictive validity was based on principals for whom several years could elapse between their initial performance rating and the benchmark estimate of their future contributions (see appendix D). Thus, the analysis required isolating a part of the measure that was stable for multiple years, which was nonexistent in the adjusted average achievement measure.

Figure 6. Relationships between the stable part of each measure and principals' future contributions to student achievement (prediction coefficients)



Note: The prediction coefficient measures the fraction of each difference in performance ratings between two principals that reflects the difference in their future contributions to student achievement. A coefficient of 1 indicates that the predictions are accurate, on average. Thus, a coefficient that is significantly different from 1 indicates inaccuracy in the performance measure. Error bars represent a 95 percent confidence interval. Figure is based on based on 268 principal-year-subject combinations and 123 distinct principals. Detailed results from these analyses are available in table F7 of appendix F.

^ Adjusted average achievement had no stable part; no differences across principals were stable for the amount of time that elapsed between principals' initial performance ratings and estimates of their subsequent contributions to student achievement.

Source: Authors' calculations based on data provided by the Pennsylvania Department of Education, 2006/7—2013/14.

Stable differences in school value-added ratings across principals partially reflected differences in principals' future contributions to student achievement, but also reflected factors outside of principals' influence. Principals with higher stable ratings on school value-added did, on average, have greater contributions to student achievement in future years. The prediction coefficient (0.43), which was significantly greater than 0, suggests that about two-fifths of each stable difference in school value-added ratings provided information about principals' future contributions to student achievement (see figure 6). However, the coefficient was also significantly less than 1, indicating inaccuracy in this measure. In other words, some of the stable differences in school value-added ratings reflected factors other than principals' contributions.

Stable differences in adjusted school value-added ratings across principals may partially predict principals' future contributions, but those predictions were very imprecise. Similar to the stable part of school value-added, the stable part of the adjusted school value-added measure had a prediction coefficient (0.54) approximately midway between zero (no accuracy) and one (ideal accuracy; see figure 6). This suggests that about half of each stable difference in adjusted school value-added ratings reflected principals' future contributions to student achievement. These results are only suggestive, but not conclusive, because the estimated prediction coefficient was quite imprecise and not statistically distinguishable from 0. In other words, despite the moderate size of the prediction coefficient, the study cannot definitively state that the stable differences in adjusted school value-added provided any information about principals' future contributions. The imprecision in the estimated prediction coefficient stemmed from the fact that there was much less variation across principals in the stable part of the performance measure than in the original measure (summary statistics on all original measures and their stable components are available in tables F1 and F2 in appendix F).

When using the original, full performance measures to predict principals' contributions to student achievement in the following year, none of the measures produced accurate predictions

Although the stable part of a performance measure is the only part that has the potential to predict principals' contributions to student achievement, using only the stable part may not be practical or desirable for states and districts. For example, focusing on the stable part of a measure does not allow for identifying improvements over time. In addition, because on some measures the stable differences in ratings are only a small fraction of the original differences in ratings, differentiation among principals on the stable part of a measure is more limited (see tables F1 and F2 in appendix F). As a result of the limited variation, most principals would receive performance ratings close to the average rating.

Given that the original measures, encompassing both the stable and unstable parts, are already implemented in states and districts and arguably more practical, the study synthesized all of the findings presented earlier to summarize the predictive validity of each original measure. This synthesis summarized the dimension of predictive validity that would be most relevant to personnel decisions: the extent to which each measure could predict principals' contributions to student achievement in the following year. Having good information about principals' contributions in the following year could be useful to states and districts making end-of-year decisions about which principals to retain, assist, or replace in the next school year.

A measure's accuracy for predicting contributions to student achievement in a subsequent year depends on both what proportion of the measure is stable (from figures 1 through 3) and how well the stable part predicts future contributions (from figure 6). Because the proportion of a measure that is stable declines over time, the accuracy of a measure's predictions will likewise decline the more years in advance the measure is used to make a prediction. Therefore, the predictive validity of a performance measure is highest when predicting contributions in the next year.

As in the earlier analysis of predictive validity, the summary of each original measure's predictive validity was also expressed as a prediction coefficient. Here, the overall prediction coefficient represented the fraction of each original difference in ratings across principals that reflected differences in their contributions in the following year.

Average achievement provided no information about principals' contributions to student achievement in the following year. When used to predict principals' contributions to student achievement in the following year, the average achievement measure had a prediction coefficient of 0.04 (figure 7). This means that average achievement was unrelated to principals' contributions in the following year. This finding was not surprising in light of earlier findings that the single-year average achievement measure was highly stable (see figures 1 and 2) but the stable part had low predictive validity (see figure 6).[7]

Average achievement provided no information about principals' contributions to student achievement in the following year.

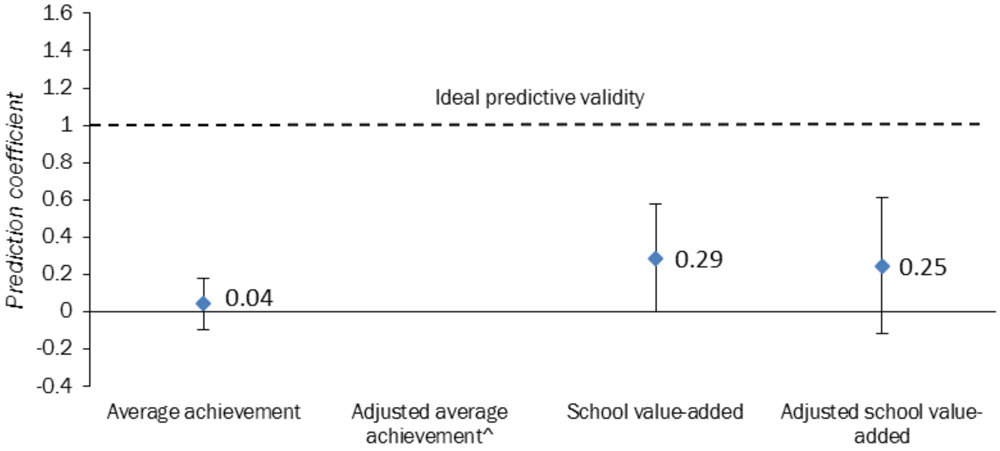
School value-added and adjusted school value-added provided, at most, a small amount of information about principals' contributions to student achievement in the following year. The prediction coefficients for both school value-added and adjusted school value-added were significantly less than 1, indicating that both measures produced inaccurate predictions of principals' contributions to student achievement in the following year (figure 7). The inaccuracy resulted from a combination of two factors. First, the stable parts of the measures were not strongly related to principals' contributions in future years (as shown earlier in figure 6). Second, both measures had a sizable unstable part that provided no information about principals' future contributions. Combined, these factors rendered the measures, at best, minimally useful for predicting contributions to student achievement in the following year.

For both value-added measures, the relationship between principals' performance ratings and their contributions to student achievement in the following year was small. As was the case with the stable part of the measure alone, principals with higher ratings on the full measure of school value-added did have greater contributions to student achievement in the following year, on average. Although the prediction coefficient (0.29) was significantly greater than 0, it was small in magnitude, suggesting that no more than one-third of each difference in school value-added ratings provided information about principals' contributions to student achievement in the following year. The prediction coefficient for adjusted school value-added (0.25) was similar in magnitude but statistically indistinguishable from zero. In comparison, analyses by Chetty et al. (2014) indicate that nearly half of each difference in teacher value-added ratings carries information about teachers' contributions in the following year (prediction coefficient of 0.49).[8]

[7] As mentioned earlier, no part of the adjusted average achievement measure was stable for the amount of time that elapsed between principals' initial ratings and estimates of their subsequent contributions. Because this measure had no stable part for which the study could assess predictive validity, the study also could not assess the entire measure's accuracy for predicting principals' contributions to student achievement in the following year.

[8] This result is based on multiplying two values provided by Chetty et al. (2014): the prediction coefficient of the stable part of teacher value-added (0.97) and the proportion of the variation in teacher value-added ratings that is stable between consecutive years (0.50). As discussed, the prediction coefficients of school value-added and adjusted school value-added are small compared to the prediction coefficient of teacher value-added. However, it is not yet known whether the prediction coefficients of the measures examined in this study are smaller or larger than those of alternative principal performance measures that do not use test scores, given that no prior research has examined the predictive validity of nontest measures.

Figure 7. Relationships between the original single-year measures and principals' contributions to student achievement in the next year (prediction coefficients)



Note: The prediction coefficient measures the fraction of each difference in performance ratings between two principals that reflects the difference in their future contributions to student achievement. A coefficient of 1 indicates that the predictions are accurate, on average. Thus, a coefficient that is significantly different from 1 indicates inaccuracy in the performance measure. Error bars represent a 95 percent confidence interval. Figure is based on a simulation that used findings on the stability of each measure (from figures 1 and 2) and the predictive validity of the stable part (from figure 6; see appendix D for details). Detailed results from these analyses are available in table F7 of appendix F.

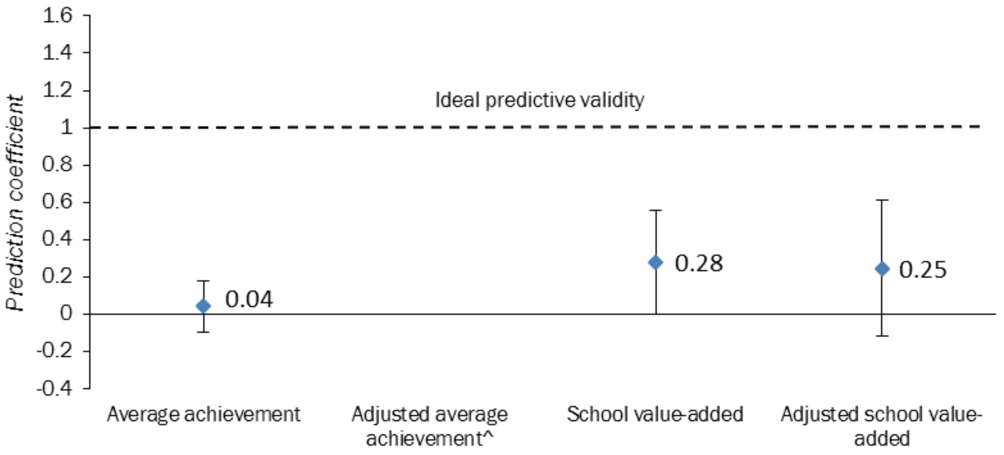
^ Adjusted average achievement had no stable part; no differences across principals were stable for the amount of time that elapsed between principals' initial performance ratings and estimates of their subsequent contributions to student achievement.

Source: Authors' calculations based on data provided by the Pennsylvania Department of Education, 2006/7—2013/14.

School value-added and adjusted school value-added provided, at most, a small amount of information about principals' contributions to student achievement in the following year.

Using a three-year rolling average of performance ratings would not improve the accuracy of the measures for predicting principals' contributions to student achievement in the next year. On each measure, three-year rolling averages of the ratings produced similarly inaccurate predictions of principals' contributions in the next year as did the single-year ratings. In fact, the prediction coefficients of the three-year averages were nearly identical to the prediction coefficients of the single-year ratings (figure 8). Because both single-year ratings and three-year averages shared the same stable part of the measure, the predictive validity of the stable part was the same across both types of ratings. Thus, the two types of ratings could differ in their accuracy only if they differed in their stability. However, as shown earlier in figure 3, the percentage of the variation across principals that was stable to the next year was approximately the same whether single-year ratings or three-year rolling averages were used. As such, the similarity of prediction coefficients between the single-year ratings and three-year averages was not surprising.

Figure 8. Relationships between three-year rolling averages of performance ratings and principals' contributions to student achievement in the next year (prediction coefficients)



Note: The prediction coefficient measures the fraction of each difference in performance ratings between two principals that reflects the difference in their future contributions to student achievement. A coefficient of 1 indicates that the predictions are accurate, on average. Thus, a coefficient that is significantly different from 1 indicates inaccuracy in the performance measure. Error bars represent a 95 percent confidence interval. Figure is based on a simulation that used findings on the stability of each measure (from figure 3) and the predictive validity of the stable part (from figure 6; see appendix D for details). Detailed results from these analyses are available in table F7 of appendix F.

^ Adjusted average achievement had no stable part; no differences across principals were stable for the amount of time that elapsed between principals' initial performance ratings and estimates of their subsequent contributions to student achievement.

Source: Authors' calculations based on data provided by the Pennsylvania Department of Education, 2006/7—2013/14.

On each measure, three-year rolling averages of the ratings produced similarly inaccurate predictions of principals' contributions in the next year as did the single-year ratings.

Implications and limitations of the study findings

To be useful for personnel decisions, a principal performance measure must have predictive validity. More specifically, a performance measure based on student achievement should successfully predict principals' contributions to student achievement in the future. The study examined the predictive validity of four test-based measures that a state or district could implement broadly in a principal evaluation system. Two of the measures, average achievement and adjusted average achievement, did not account for students' past achievement. Two other measures, school value-added and adjusted school value-added, accounted for students' past achievement by measuring students' achievement growth.

None of the four performance measures provided good predictions about principals' contributions to student achievement in the following year. Average achievement and adjusted average achievement were entirely unrelated to principals' contributions in the following year. School value-added and adjusted school value-added were, at best, weakly related to principals' contributions in the following year. For the two value-added measures, less than one-third of each difference in performance ratings across principals reflected differences in their contributions in the subsequent year.

To understand why these performance measures did not accurately predict principals' future contributions to student achievement, this study considered two key factors that shape a measure's predictive validity. First, performance measures that are more stable have greater potential for predictive validity. Only the stable part of a measure—the part representing persistent differences across principals in their ratings—can provide any information at all about principals' future performance. However, stable ratings could still be uninformative if they capture persistent factors other than principals' contributions to student achievement. Therefore, the second key condition for a measure to have high predictive validity is that the stable part of the measure must be strongly associated with principals' future contributions.

Each of the four measures failed to demonstrate high predictive validity for a different combination of reasons. Average achievement was very stable across years, but the stability reflected influences on student achievement that were outside of principals' control. Adjusted average achievement had very low stability—with no differences in ratings remaining stable for more than one year—so it reflected transient factors that were uninformative about how well principals would perform in the future. Both school value-added and adjusted school value-added produced ratings that were only partially stable, and the stable ratings from both measures only partially reflected principals' contributions. The combination of those factors led the value-added measures to yield, at most, a small amount of information for predicting principals' contributions in the following year.

Given the modest stability of the value-added measures, increasing their stability could conceivably have enhanced their predictive validity. Although some evaluation systems average ratings across multiple years under the assumption that doing so enhances the stability of a performance measure, this study did not find evidence to support this claim. For both value-added measures, a three-year average of ratings had about the same degree of persistence to the following year as the most recent single-year rating did. Although some of the unstable parts of the most recent year's rating were offset when averaged with additional prior years of ratings, those additional prior ratings also reflected some outdated aspects of a principal's performance that would not persist to the future. On net, those two opposite influences led to no change in stability—and, therefore, no improvement in predictive validity—when using three years instead of one year of ratings from the value-added measures.

Suggestions for the design of principal evaluation systems and future research

Based on this study and other existing research, no available measures of principal performance have yet been shown to accurately identify principals who will contribute successfully to student outcomes in future years. Until now, no existing research had ever examined the predictive validity of most test-based measures considered in this study, with the exception of school value-added, which Chiang et al. (2016) found to produce inaccurate predictions of principals' future contributions. The current study found little evidence that any widely feasible test-based measures could accurately predict principals' contributions in the following year.

At the same time, nontest measures of principal performance, such as ratings of leadership practices and surveys of teachers, currently have no better evidence to support their validity than do the test-based measures. Most of these nontest measures have never been examined for their validity (see McCullough et al. [2016] for a review of the evidence). Of particular importance, no research has ever determined whether such measures predict principals' future contributions to student outcomes.

The lack of evidence to support the validity of existing principal performance measures underscores the need to gather more evidence in search of accurate measures. In the meantime, states and districts may still need to evaluate principals in some manner. The remainder of this section discusses the study's implications for the design of principal evaluation systems and the direction of future research.

When evaluating principals, emphasize measures that provide at least some information about principals' future contributions to student outcomes. As discussed throughout this report, many of the reasons for evaluating principals stem from wanting to know how effectively they will contribute to student outcomes in future years (recognizing that test scores may not be the only outcomes of interest). No test-based measure has met the bar of predicting principals' future contributions accurately. Nevertheless, if states and districts need to choose measures, they ought to prefer those that are at least somewhat associated with principals' future contributions, over those that are completely unrelated to principals' future contributions.

Based on the evidence from this study, the two measures that did not account for students' past achievement—average achievement and adjusted average achievement—should not be used in a principal evaluation system. Those measures provided no information about principals' future contributions to student achievement. As a consequence, some measures that are common in school accountability systems are likely ill-suited to evaluating principals. A school's proficiency rate—the percentage of students who score proficient or above—is conceptually similar to average achievement; the change in a school's proficiency rate between two years is conceptually similar to adjusted average achievement. Using those measures to evaluate principals could lead to very inaccurate conclusions about which principals will be effective or ineffective at contributing to student achievement in subsequent years.

Instead, principal evaluation systems that use test-based measures should emphasize ratings based on school value-added or adjusted school value-added. Those are the only two test-based measures shown to have any degree of predictive validity. In particular, this study found that the value-added measures may provide some—although not much—information for predicting principals' contributions in the following year. School value-added had a small but statistically significant relationship with principals' contributions in the following year. The relationship between adjusted school value-added and principals' contributions in the following year was not statistically significant, but the magnitude was similar to that which the study found for school value-added. Therefore, using the value-added measures leads to better predictions of principals' future contributions than not using those measures, but the measures will still make plenty of mistakes when trying to identify principals who will contribute effectively or ineffectively to student achievement in future years.

As discussed earlier, nontest measures have not yet been examined for their predictive validity. Therefore, states and districts lack a clear basis for deciding whether and how to use such measures in principal evaluations. For states and districts seeking to include nontest measures in an evaluation system, a possible approach is to select nontest measures that are associated with test-based measures containing at least some predictive validity. For example, prior research has shown that Pennsylvania's measure of principals' leadership practices is associated with adjusted school value-added (McCullough et al., 2016). Unfortunately, few other nontest measures of principal performance are supported by this type of validity evidence. In fact, few nontest measures have ever been shown to meet even the basic conditions necessary for being valid, such as demonstrating adequate reliability (Goldring, Cravens, Murphy, Porter, Elliott, & Carson, 2009; Condon & Clifford, 2012). Therefore, as a first step toward implementing nontest measures that have any potential for predicting principals' contributions, states and districts should work to ensure that these measures are grounded in best practices for achieving high reliability, such as extensive training of raters (for ratings of leadership practices) and high response rates (for surveys of teachers).

Exercise caution when using existing test-based measures to make major decisions about principals. If states and districts decide to evaluate principals with test-based measures, they will face choices about what types of personnel decisions the measures will influence and how closely those decisions will be tied to the measures. Personnel decisions range from relatively minor, such as selecting principals to provide or receive extra mentoring, to very consequential, such as identifying which principals will receive large bonuses or be subject to dismissal. As discussed earlier, although the value-added measures contain some information about principals' future contributions, those measures still have a great deal of inaccuracy and could lead to many erroneous personnel decisions. The more consequential the decision, the greater the potential harm to principals' careers that might stem from erroneous decisions—and, therefore, the greater the need for caution in using test-based measures to make those decisions. States and districts will need to exercise their own, careful judgments about their willingness to tolerate error in each type of personnel decision before deciding whether and how closely to tie it to test-based performance ratings.

Gather more evidence from large states with longer data histories and examine more measures. It is important to know whether the study's conclusions about the low predictive validity of test-based performance measures apply to principals outside of this study. The study's method for validating principal performance measures, an adaptation of the method Chetty et al. (2014) used to validate teacher value-added measures, greatly limited the number of principals who could be included in the study sample. Principals needed to be involved in a school leadership transition where it was possible not only to measure the contributions to student achievement of both the departing and the incoming principal but also to measure each principal's performance at specific other points in time. Although the study team had access to seven school years of data on all students and principals in Pennsylvania, only 123 principals could be included in the prediction analysis. Future studies using this method should consider using data from several large states that have even longer data histories, given that the number of schools and school years are key factors influencing the size of the study sample.

Although the four measures in this study represent the currently available approaches to evaluating principals based on student test scores, researchers could explore refinements to these measures with the aim of enhancing predictive validity. For several measures, instability contributed to their low predictive validity, so researchers could try to identify ways to increase stability. Although this study found that averaging across multiple years did not improve stability, there may be other strategies to consider, such as adjusting the measures for longer histories of students' past achievement and schools' past ratings.

Given the inaccuracy of the test-based measures, state and district leaders and researchers should also make every effort to identify nontest measures that can predict principals' future contributions to student outcomes. To date, no studies have examined the predictive validity of nontest measures. This study's benchmark approach to measuring principals' contributions based on school leadership transitions could be readily applied to assess the predictive validity of nontest measures.

Limitations of the study

As mentioned earlier, only a small sample of departing and incoming principals could be included in the analysis of predictive validity. These principals and the schools they led differed in some regards from the state as a whole (see appendix E). Incoming principals were less experienced than principals overall and more likely to be racial/ethnic minorities and women. Departing principals had the opposite characteristics. The schools these principals led had larger proportions of students receiving free or reduced-price lunches and who are black or Hispanic, on average, compared with Pennsylvania students overall. These differences also raise the possibility that principals and schools in the study might have differed from the rest of the state in unmeasured ways as well. Therefore, findings might differ for principals who were not in the study.

All measures of principal performance in this study were based on student outcome measures from state tests. This study did not examine the predictive validity of nontest measures—such as school climate measures, professional practice measures, or teacher retention measures—which might also provide important information about the quality of principals' leadership.

Finally, there remains no clear consensus on the best method of measuring principals' contributions to student achievement. The study's benchmark approach to measuring differences in contributions between departing and incoming principals was to calculate the change in student achievement from the departing principal's final year to the incoming principal's second year. Although this study did not find clear threats to the validity of this benchmark approach (see appendix F), there remained the possibility that changes in achievement during leadership transitions could have been driven by other unmeasured factors, leading to an inaccurate measure of the principals' contributions. If the benchmark results were inaccurate, then the study findings did not provide a test of the predictive validity of principal performance measures.

Appendix A. Data used in the study

The Pennsylvania Department of Education supplied all the data for this study. Specifically, the department authorized the study team to access database records on all students and principals in Pennsylvania that two of its agencies—the Bureau of Assessment and Accountability and the Division of Data Quality—maintain. The key information included student achievement data, student demographic data, and principal data.

Student achievement data

The study team used achievement scores for all Pennsylvania students in grades 3–8 who were administered end-of-grade state assessments from 2006/07 to 2013/14 (eight years of test scores). The assessment, called the Pennsylvania System of School Assessment (PSSA), is administered annually in reading and math in grades 3–8; science in grades 4 and 8; and writing in grades 5 and 8. The student assessment data included scores from modified PSSA tests that students with disabilities are eligible to take, based on their individualized education program. The study team standardized all test scores to have a 0 mean and a standard deviation equal to 1 within each subject, grade, and year.

Student demographic data

The study team used student demographic information from the state's longitudinal education data system, called the Pennsylvania Information Management System (PIMS). PIMS includes records on all students in the state's public schools from 2007/08 to 2013/14 (seven years of student data). Every student is assigned a unique identification number that is consistent across years. For each student in each year, the data indicate the schools the student attended and the student's gender, age, race/ethnicity, free and reduced-price lunch status, English learner student status, and special education status.

Principal data

The study team also used PIMS information on school principals from 2007/08 to 2013/14 (seven years of principal data). The data on principals indicated the schools to which they were assigned as well as their gender, education degrees, race/ethnicity, and total work experience in PreK–12 education. This study focused on principals who led schools that served students in any grades 3–8.

Appendix B. Technical details of the principal performance measures that the study assessed

This appendix describes the technical approach for calculating principal performance ratings on each performance measure examined in this study. These measures represented different approaches to estimating principals' contributions to student achievement. Although each measure was intended to estimate principals' contributions to student achievement, it may or may not have been a valid method for doing so. The study's purpose was to assess each measure's validity based on the methods described in appendix D—specifically, by assessing how well performance ratings from one period (known as the measurement period) could predict principals' contributions in a different period (known as the validation period).

To determine which principal performance measures should be assessed, the study used three criteria for inclusion that approximate the criteria that a district or state might use to select evaluation measures. Each eligible performance measure for the study was based on student achievement data, could measure the performance of most principals in a given year, and could compare the performance of each principal to that of any other principal.

The study assessed the validity of the following measures, each of which met the three criteria:

- **Average achievement:** average student score on the state assessment—the Pennsylvania System of School Assessment (PSSA)—across all tested students in a principal's school.
- **Adjusted average achievement:** average achievement, adjusted for prior average achievement at the school.
- **School value-added:** difference between students' achievement growth and the growth of other students in the state with similar prior scores and other characteristics.
- **Adjusted school value-added:** school value-added, adjusted for the school's prior value-added.

Because adjusted school value-added is the most complex measure, this appendix first describes all of the steps needed to construct this measure. It then explains how each of the other measures can be adapted from a subset of the steps used in constructing adjusted school value-added.

Adjusted school value-added

Adjusted school value-added was the difference between the current value-added of a principal's school and what it was predicted to be based on the school's value-added from a prior year. Because the prediction was determined by the school's prior value-added, it was supposed to encapsulate persistent factors at the school, such as neighborhood quality, that influenced student achievement growth. Subtracting this prediction essentially insulated principals from being held accountable for persistent school characteristics beyond their control.

To estimate adjusted school value-added, the study began by estimating school value-added, separately by grade, year, and subject. The study estimated schools' contributions to students' one-year achievement growth (one-year value-added) and growth over two consecutive years (two-year value-added). Next, value-added ratings were averaged across grades, within each subject-year combination. Finally, value-added ratings were adjusted to account for prior school value-added and were transformed into empirical Bayes estimates to account for measurement error. The remainder of this section describes each of these steps in detail.

The study focused primarily on two-year, rather than one-year, adjusted school value-added measures because the benchmark estimates from the validation period—the key point of comparison—also reflected principals' two-year contributions (see appendix D). Despite capturing students' two-year growth, adjusted school value-added was still considered a single-year performance measure because all of the data on students' final achievement was based on a single year. This measure compared students' final achievement to predictions based on their own scores from two years earlier and the schools' value-added from two years earlier, but nevertheless the final scores came from one year only.

Initial value-added estimates by grade, year, and subject. The first step was to estimate school value-added separately by grade, year, and subject.

Two-year school value-added estimates were based on PSSA scores in the following subjects, grades, and school years, separately for each combination:

- PSSA math: grades 5, 6, 7, and 8 (2009/10–2013/14)
- PSSA reading: grades 5, 6, 7, and 8 (2009/10–2013/14)

Using student outcome data for each year, the study estimated a value-added model (VAM) of the following form, separately by grade and subject:

$$\text{(B1)} \qquad A_{igstk} = \beta' P_{i(t-2)} + \gamma' X_{igstk} + \delta' S_{igstk} + e_{igstk}$$

In the model, A_{igstk} was the assessment score for student *i* in grade *g* attending school *s* in year *t* for subject *k*, expressed as a *z*-score with mean 0 and standard deviation 1 within each subject-grade-year combination. The vector $P_{i(t-2)}$ included variables for student *i*'s PSSA scores in math and reading two years before year *t*. The vector X_{igstk} was a set of variables for observed student characteristics, including free and reduced-price lunch participation, race/ethnicity, English learner student status, disability status, gender, age, mobility, grade repetition, and test modification—that is, whether the assessment score was a PSSA-Modified score. The VAMs included PSSA-Modified scores as the outcomes for students with disabilities who were eligible to take modified assessments as a result of their individualized education program. The coefficients 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_{igstk} was a random error term.

The vector S_{igstk} included an indicator for each school in the VAM that was equal to 1 for students attending the school and 0 otherwise. Students attending multiple schools were included in the model on multiple rows of the dataset, once for each school, and each student-school-year observation had exactly one nonzero element in S_{igstk}. Weights were used to account for a student's exposure to each school that he or she attended during the school year. In each 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.

All students with a baseline test score in the same subject were included in the VAMs. Students' other baseline scores and baseline characteristics were imputed if they were missing. Only a small fraction (fewer than 1 percent) of values were imputed for any baseline test score or baseline characteristic. The study used regression models to predict values of missing baseline scores and baseline characteristics based on other prior-year scores, outcome scores, and background characteristics with non-missing values. Specifically, for any student-year row with a missing baseline score or characteristic, the study estimated a regression of the baseline score or characteristic on the set of prior-year scores, outcome scores, and background characteristics for which the row had non-missing values, using all student-year rows in the dataset with non-missing values for those specific covariates. The study then predicted a value for the missing baseline score or characteristic based on the estimated regression.

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

The errors-in-variables regression approach entailed a two-step process necessary to facilitate the accurate calculation of standard errors. First, equation B1 was estimated separately for each grade-subject-year combination 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 the baseline math and reading scores:

(B2) $$\tilde{A}_{igstk} = A_{igstk} - \beta' P_{i(t-2)} .$$

The second step was to use the adjusted outcome, $\tilde{A}_{igstk}$, in place of the actual score and estimate equation B3 by ordinary least squares regression, separately for each grade-subject-year:

(B3) $$\tilde{A}_{igstk} = \gamma' X_{igstk} + \delta' S_{igstk} + e_{igstk} .$$

The student-level residuals from equation B3 were outputted:

(B4) $$u_{igstk} = \tilde{A}_{igstk} - \hat{\gamma}' X_{igstk} .$$

Finally, weighting for the student's exposure to the school described previously $\left(w_{igst} \right)$, the study calculated each school's estimated contribution to student achievement growth over a two-year period $\left(E_{gstk} \right)$ as

(B5) $$E_{gstk} = \frac{\sum_i u_{igstk} w_{igst}}{\sum_i w_{igst}} .$$

The study implemented a nearly identical approach to estimate schools' contributions to student achievement growth over a one-year period. One-year school value-added estimates were based on PSSA scores in the following subjects, grades, and school years, separately for each combination:

- PSSA math: grades 4, 5, 6, 7, and 8 (2007/08–2013/14)
- PSSA reading: grades 4, 5, 6, 7, and 8 (2007/08–2013/14)

One-year school VAMs were estimated in the following form, separately by grade and subject for each outcome year:

(B6) $$A_{igstk} = \beta' P_{i(t-1)} + \gamma' X_{igstk} + \delta' S_{igstk} + e_{igs}$$

In the model, the vector $P_{i(t-1)}$ included student *i*'s PSSA math and reading scores one year before year *t*, and all other variables were defined in the same way as they were in equation B1. Using this framework, the study then followed the same steps used in the two-year school value-added estimates to estimate one-year school contributions to student growth.

Estimating variances and covariances of the estimation errors. All value-added estimates contained some degree of error. To prepare to adjust the estimates to account for this measurement error, the variances and covariances of the estimation errors for each value-added estimate were calculated.

For the value-added estimate, the variance of the estimation error was calculated as

$$\text{(B7)}\qquad V^{error}_{gstk} \equiv Var\left(E_{gstk}\right) = \left(V^{residual}_{gtk}\right)\left(Q_{gstk}\right)/\left(W^{2}_{gstk}\right),$$

where $V^{residual}_{gtk}$ was the average within-school variance of the residuals $\left(u_{igstk}\right)$ for grade *g* and subject *k* in year *t*; Q_{gstk} was the sum of squared weights for students in school *s* who had a test score in grade *g*, year *t*, and subject *k*; and W_{gstk} was the total weight for students in school *s* who had a test score in grade *g*, year *t*, and subject *k*.

The study also estimated the covariance between the estimation errors of each school's math and reading value-added estimates in the same year and grade, as follows:

$$\text{(B8)}\qquad C^{error}_{gst} \equiv Cov\left(E_{gst,math}, E_{gst,read}\right) = \frac{\left(C^{residual}_{gt}\right)\left(Q^{both}_{gst}\right)}{\left(W_{gst,math}\right)\left(W_{gst,read}\right)},$$

where $C^{residual}_{gt}$ was the average within-school covariance of the math and reading residuals for grade *g* in year *t*, and Q^{both}_{gst} was the sum of squared weights for students in school *s* who had both math and reading test scores in grade *g* and year *t*.

Averaging across grades. To generate a single math value-added estimate and a single reading value-added estimate for each school in a given year, the study calculated a weighted average across grades in a school, with grades weighted by the effective number of students (that is, the total student weight in the grade, W_{gstk}) at the school:

$$\text{(B9)}\qquad E_{stk} = \left(\sum\nolimits_{g\in G_{st}} W_{gstk}E_{gstk}\right)/\left(\sum\nolimits_{g\in G_{st}} W_{gstk}\right).$$

The subset of grades included in the average, G_{st}, varied based on whether the estimates were for two-year contributions to student growth or one-year contributions to student growth. Specifically, the average one-year school value-added estimate for each school-year combination included all grades 4–8 or the subset of all grades 4–8 in which students were enrolled and tested at the school. The average two-year school value-added estimate for each school-year combination included all grades 5–8 or the subset of all grades 5–8 in which students were enrolled and tested at the school, with the exception of the entry grade. That is, if a school's entry grade was any grade from 5–8, the entry grade was not included in the average two-year value-added estimate. The entry grade was dropped because the two-year value-added estimate was intended to reflect the contribution of a school to a student's growth over a two-year period. Students enrolled in the entry grade of a school in the year of interest would have had only one year of exposure to a school at the time of testing.

The variance of the estimation error for each aggregate estimate (B10) and the covariance of the estimation errors for the aggregate math and aggregate reading estimates (B11) were also calculated for each school-year combination. The same weights and specific grade subsets used to aggregate the value-added estimates were used to calculate the variances and covariances of the estimation errors.

$$\text{(B10)} \quad V_{stk}^{error} \equiv Var\left(E_{stk}\right)=\left(\sum\nolimits_{g\in G_{st}} W_{gstk}^{2} V_{gstk}^{error}\right)/\left(\sum\nolimits_{g\in G_{st}} W_{gstk}\right)^{2}$$

$$\text{(B11)} \quad C_{st}^{error} \equiv Cov\left(E_{st,math}, E_{st,read}\right)=\frac{\left(\sum\nolimits_{g\in G_{st}} W_{gst,math} W_{gst,read} C_{gst}^{error}\right)}{\left(\sum\nolimits_{g\in G_{st}} W_{gst,math}\right)\left(\sum\nolimits_{g\in G_{st}} W_{gst,read}\right)}$$

Obtaining best linear predictors of school value-added. The original school value-added estimates that the study calculated in equation (B9) were not the best predictors of true school value-added. Because the initial estimates were based on a finite number of students, they were subject to some random estimation error. The study corrected for this error by adjusting each estimate using an empirical Bayes "shrinkage" procedure, which effectively "shrank" the estimate toward the average estimate in the study sample by a factor dependent on the precision of the initial estimate. For schools with relatively imprecise original 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 original 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. The shrinkage procedure also accounted for the information available about school value-added in a different subject; that is, a school's value-added estimate in reading provided some additional information to predict the school's true value-added in math, and vice versa.

The shrunk school value-added estimate, or best linear predictor of school value-added, in any subject-year combination was calculated as the weighted sum of the original school value-added estimates in math and reading, with weights reflecting shrinkage as well as how strongly or weakly success in one subject predicted success in another.

To construct these estimates, the first step was to take the sample variance of the initial value-added estimates in a given subject-year combination and subtract the average squared standard error of those estimates. This difference ($\sigma_{t,read}^{2}$ for reading and $\sigma_{t,math}^{2}$ for math) measured the variance of true school value-added—that is, the variance that would occur if school value-added were observed without error. Similarly, the study took the sample covariance between the math and reading value-added estimates in the same year and subtracted the average covariance of the errors of those estimates. This difference (θ_{t})

measured the covariance between true school value-added in math and reading—that is, the covariance that would occur if school value-added were observed without error. Next, two sets of shrinkage weights were constructed. The first set, for shrunk math estimates, was constructed using the formulas in equations B12a and B12b. The second set, for shrunk reading estimates, was constructed using the formulas in equations B13a and B13b.

For every shrunk math estimate, the shrinkage weight on the original math value-added estimate (equation B12a) and the shrinkage weight on the original reading value-added estimate (equation B12b) were constructed as

$$\text{(B12a)}\qquad \gamma_{st,math}^{math}=\frac{\left(\sigma_{t,read}^{2}+V_{st,read}^{error}\right)\sigma_{t,math}^{2}-\left(\theta_{t}+C_{st}^{error}\right)\theta_{t}}{\left(\sigma_{t,math}^{2}+V_{st,math}^{error}\right)\left(\sigma_{t,read}^{2}+V_{st,read}^{error}\right)-\left(\theta_{t}+C_{st}^{error}\right)^{2}}$$

$$\text{(B12b)}\qquad \gamma_{st,read}^{math}=\frac{\left(\sigma_{t,math}^{2}+V_{st,math}^{error}\right)\theta_{t}-\sigma_{t,math}^{2}\left(\theta_{t}+C_{st}^{error}\right)}{\left(\sigma_{t,math}^{2}+V_{st,math}^{error}\right)\left(\sigma_{t,read}^{2}+V_{st,read}^{error}\right)-\left(\theta_{t}+C_{st}^{error}\right)^{2}}$$

where $\sigma_{t,read}^{2}$ was the variance of true school value-added in reading in year *t*, $\sigma_{t,math}^{2}$ was the variance of true school value-added in math in year *t*, θ_{t} was the covariance between true school value-added in reading and true school value-added in math, $V_{st,read}^{error}$ was the variance of the estimation error for the original reading value-added estimate, $V_{st,math}^{error}$ was the variance of the estimation error for the original math value-added estimate, and C_{st}^{error} was the covariance between the estimation errors of the original math and reading value-added estimates for school *s* in year *t*.

Similarly, for every shrunk reading estimate, the shrinkage weight on the original reading value-added estimate (equation B13a) and the shrinkage weight on the original math value-added estimate (equation B13b) were constructed as:

$$\text{(B13a)}\qquad \gamma_{st,read}^{read}=\frac{\left(\sigma_{t,math}^{2}+V_{st,math}^{error}\right)\sigma_{t,read}^{2}-\left(\theta_{t}+C_{st}^{error}\right)\theta_{t}}{\left(\sigma_{t,math}^{2}+V_{st,math}^{error}\right)\left(\sigma_{t,read}^{2}+V_{st,read}^{error}\right)-\left(\theta_{t}+C_{st}^{error}\right)^{2}}$$

$$\text{(B13b)}\qquad \gamma_{st,math}^{read}=\frac{\left(\sigma_{t,read}^{2}+V_{st,read}^{error}\right)\theta_{t}-\sigma_{t,read}^{2}\left(\theta_{t}+C_{st}^{error}\right)}{\left(\sigma_{t,math}^{2}+V_{st,math}^{error}\right)\left(\sigma_{t,read}^{2}+V_{st,read}^{error}\right)-\left(\theta_{t}+C_{st}^{error}\right)^{2}}$$

After constructing these weights, the study generated shrunk school value-added estimates for math ($E_{st,math}^{shrunk}$) that were equal to the weighted sum of the original math value-added estimate and the original reading value-added estimate, where the weights were the quantities generated in equations B12a and B12b, respectively:

(B14) $E_{st,math}^{shrunk} = \gamma_{st,math}^{math} E_{st,math} + \gamma_{st,read}^{math} E_{st,read}$.

The variance of the estimation error for the shrunk estimate was calculated as

(B15) $VS_{st,math}^{error} \equiv Var\left(E_{st,math}^{shrunk}\right)$

$$= \left(\gamma_{st,math}^{math}\right)^2 V_{st,math}^{error} + \left(\gamma_{st,read}^{math}\right)^2 V_{st,read}^{error} + 2\gamma_{st,math}^{math}\gamma_{st,read}^{math} C_{st}^{error} .$$

Likewise, the study generated shrunk school value-added estimates for reading $\left(E_{st,read}^{shrunk}\right)$ that were equal to the weighted sum of the original reading value-added estimate and the original math value-added estimate, in which the weights were the quantities generated in equations B13a and B13b:

(B16) $E_{st,read}^{shrunk} = \gamma_{st,read}^{read} E_{st,read} + \gamma_{st,math}^{read} E_{st,math}$.

The variance of the estimation error for $E_{st,read}^{shrunk}$ was calculated as

(B17) $VS_{st,read}^{error} \equiv Var\left(E_{st,read}^{shrunk}\right)$

$$= \left(\gamma_{st,read}^{read}\right)^2 V_{st,read}^{error} + \left(\gamma_{st,math}^{read}\right)^2 V_{st,math}^{error} + 2\gamma_{st,read}^{read}\gamma_{st,math}^{read} C_{st}^{error} .$$

Adjusting for prior school value-added. The study adjusted the current school value-added estimates based on the school's prior school value-added to better isolate the contribution of the principal. Current school value-added estimates might have reflected many factors beyond the current principal's leadership that influenced student outcomes, including the continued effects of previous principals and persistent school characteristics. Adjustment for the school's prior value-added was aimed at removing the effects of these persistent or lingering factors that were beyond the current principals' control.

The adjustment was based on the following regression model:

(B18) $E_{spt} = \beta_0 + \beta_1 E_{s(t-2)}^{shrunk} + v_{spt}$,

where E_{spt} was the unshrunk estimate of the current value-added of school *s* led by principal *p* in year *t*, and $E_{s(t-2)}^{shrunk}$ was the shrunk value-added estimate of the same school from two years before that year. With the true coefficients β_0 and β_1, the quantity $\left(\beta_0 + \beta_1 E_{s(t-2)}^{shrunk}\right)$ was defined as the expected value-added a school would demonstrate in year *t* if it started at a value-added of $E_{s(t-2)}^{shrunk}$ two years ago and was subsequently led by the average principal in the state. The error term, v_{spt} , represented all other factors—including principals' contributions—that led a school to deviate from its expected value-added. After estimating this equation by ordinary least squares (OLS), the study constructed a prediction, $\left(\hat{\beta}_0 + \hat{\beta}_1 E_{s(t-2)}^{shrunk}\right)$, for the current value-added of every school. The adjusted school value-added estimate of principal *p* was the difference between the school's actual and predicted value-added:

$$\text{(B19)} \qquad \hat{v}_{spt} = E_{st} - \left(\hat{\beta}_0 + \hat{\beta}_1 E_{s(t-2)}^{shrunk}\right).$$

As with any value-added measure, the adjusted school value-added measure was intended to measure relative performance. Because the prediction, $\left(\hat{\beta}_0 + \hat{\beta}_1 E_{s(t-2)}^{shrunk}\right)$, was intended to capture a school's value-added under the average principal, each adjusted school value-added estimate measured a principal's effectiveness relative to the average principal. Specifically, it estimated how much different a student's achievement would be from having a specific principal for two years compared with having the average principal. As discussed earlier, performance measures in the measurement period focused on principals' two-year contributions to be consistent with estimates in the validation period, which also captured principals' two-year contributions.

In each year, the study generated adjusted school value-added estimates for all principals in Pennsylvania who were in their second or later year at their school, and whose current school had estimates of school value-added in both the current year and two years before that year. Although current school value-added estimates were two-year estimates, the baseline school value-added estimates were one-year estimates to minimize the number of years of data required for this performance measure. One-year estimates (or, in fact, any baseline variables) were appropriate as the independent variable in equation B18 as long as they were strongly predictive of current school value-added.

In equation B18, the baseline school value-added estimate $\left(E_{s(t-2)}^{shrunk}\right)$ was a shrunk estimate (calculated using the procedure described previously), whereas the current school value-added estimate $\left(E_{spt}\right)$ was an original, unshrunk estimate. Shrinking the baseline school value-added estimates effectively reduced the measurement error in the independent variable in equation B18. The measurement error in the unshrunk estimates would have attenuated the estimated coefficient $\left(\hat{\beta}_1\right)$; using shrunk estimates eliminated this attenuation bias. In contrast, measurement error in the dependent variable, or current school value-added estimate, does not bias the coefficient estimates. In fact, using shrunk estimates for current school value-added would have inappropriately compressed the variation in the dependent variable and thereby bias downward the estimated coefficient.

For predictions from equation B18, $\left(\hat{\beta}_0 + \hat{\beta}_1 E_{s(t-2)}^{shrunk}\right)$, to represent how schools would have performed under the average principal, principals' contributions to student achievement must not, on average, vary with schools' prior value-added. For example, if principals with large contributions were disproportionately working in schools with high prior value-added, the prediction of these schools' current value-added would be too high. The reason is that the prediction of these schools' current value-added would essentially be based on how well schools with similarly high prior value-added were performing in the current year—and these schools would be disproportionately staffed by principals with large contributions, not the average principal.

Equation B18 can illustrate the importance of assuming that mean principal effectiveness must not vary with schools' prior value-added. The true coefficient, β_1, represented the relationship between prior and current school value added if the average principal were to have led the school in the intervening period. The OLS estimate for this coefficient had an expected value of

$$\text{(B20)} \quad E\left(\hat{\beta}_1\right)=\frac{Cov\left(E_{spt},E_{s(t-2)}^{shrunk}\right)}{Var\left(E_{s(t-2)}^{shrunk}\right)}=\frac{Cov\left(\beta_0+\beta_1 E_{s(t-2)}^{shrunk}+v_{spt},E_{s(t-2)}^{shrunk}\right)}{Var\left(E_{s(t-2)}^{shrunk}\right)}$$

$$=\beta_1+\frac{Cov\left(v_{spt},E_{s(t-2)}^{shrunk}\right)}{Var\left(E_{s(t-2)}^{shrunk}\right)}.$$

Therefore, the estimate $\hat{\beta}_1$ was unbiased only if $Cov\left(v_{spt},E_{s(t-2)}^{shrunk}\right)=0$—that is, if principals' true contributions were uncorrelated with the prior value-added of the schools they lead. If principals with large contributions were disproportionately likely to lead schools with high prior value-added—such that $Cov\left(v_{spt},E_{s(t-2)}^{shrunk}\right)>0$—then $\hat{\beta}_1$ would be systematically overestimated, causing schools with high prior value-added to have current-year predictions that were too high.

The upshot is that it was critically important to estimate equation B18 on a sample of principals in which principals' true contributions were not correlated with their schools' prior value-added. There was a clearly defined group of principals whose true contributions were likely to be correlated with their schools' prior value-added: principals who were already leading their schools two years ago. Within that group, principals with large contributions might have boosted their schools' prior value-added, so principals with large contributions were expected to be disproportionately represented in schools with high prior value-added. Likewise, principals with small contributions were expected to be disproportionately represented in schools with low prior value-added. To avoid this problem, the study first estimated equation B18 only on the sample of schools whose baseline and current school value-added were measured under different principals—that is, schools in which the current principal started in year (t – 1). From this regression, the study obtained estimated coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$, which were then used to generate a prediction of current school value-added for every school (both inside and outside the regression sample).

The estimation of the coefficients in equation B18 using only schools in which a principal transition occurred was advantageous because new principals could not have affected school value-added before their placement at the school. As such, this sample restriction eliminated the most obvious source of correlation between baseline school value-added and principals' contributions to student growth. However, the trajectory of schools' value-added in the first two years under new principals might differ from the trajectory of schools' value-added under the average principal in the state—including principals with long tenures at their schools. For example, new principals might face different opportunities and constraints for shaping their schools' value-added relative to veteran principals. To account for this potential difference in trajectories, the study also estimated coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ using the full sample of schools as a sensitivity analysis. Both the coefficients estimated using the full sample (left column) and the coefficients estimated using the restricted second-year principal sample (right column) are reported in table B1.

Table B1. Estimated relationships between school value-added and the same measure obtained two years earlier

Subject	Estimated relationship using all schools	Estimated relationship using schools whose principals were in their second year
Math	0.66* (0.025) [10,332]	0.47* (0.061) [925]
Reading	0.56* (0.025) [10,332]	0.40* (0.063) [925]

* Significant at *p* = 0.05.

Note: Standard errors are indicated in parentheses. Number of school-year combinations are indicated in brackets. Each coefficient estimate is derived from a separate regression in which the dependent variable is a school's unshrunk performance estimate and the independent variable is the school's shrunk performance estimate in the same subject from two years earlier. Every regression also controls for year indicators. Outcome years include 2009/10–2013/14.

Source: Authors' calculations based on data provided by the Pennsylvania Department of Education, 2006/7—2013/14.

Obtaining best linear predictors of adjusted school value-added. The adjusted school value-added estimates were then shrunk by applying the same procedure used to shrink the unadjusted school value-added estimates.

First, the average squared standard error of the adjusted value-added estimates (assumed to be identical to that of the unadjusted value-added estimates) was subtracted from the sample variance of the adjusted value-added estimates in each subject-year combination. This difference ($\sigma_{t,read}^{2\ adj}$ for reading and $\sigma_{t,math}^{2\ adj}$ for math) measured the variance of true adjusted school value-added. Similarly, the average covariance of the errors of the adjusted math and reading value-added estimates (assumed to be identical to that of the unadjusted value-added estimates) was subtracted from the sample covariance between the math and reading value-added estimates in the same year. This difference $\left(\theta_t^{adj}\right)$ measured the covariance between true adjusted school value-added in math and reading.

Shrinkage weights for the adjusted math estimates were then constructed as

$$\text{(B21a)}\quad \gamma_{st,math}^{adj\,math} = \frac{\left(\sigma_{t,read}^{2\,adj} + V_{st,read}^{error}\right)\sigma_{t,math}^{2\,adj} - \left(\theta_t^{adj} + C_{st}^{error}\right)\theta_t^{adj}}{\left(\sigma_{t,math}^{2\,adj} + V_{st,math}^{error}\right)\left(\sigma_{t,read}^{2\,adj} + V_{st,read}^{error}\right) - \left(\theta_t^{adj} + C_{st}^{error}\right)^2} \quad \text{and}$$

$$\text{(B21b)}\quad \gamma_{st,read}^{adj\,math} = \frac{\left(\sigma_{t,math}^{2\,adj} + V_{st,math}^{error}\right)\theta_t^{adj} - \sigma_{t,math}^{2\,adj}\left(\theta_t^{adj} + C_{st}^{error}\right)}{\left(\sigma_{t,math}^{2\,adj} + V_{st,math}^{error}\right)\left(\sigma_{t,read}^{2\,adj} + V_{st,read}^{error}\right) - \left(\theta_t^{adj} + C_{st}^{error}\right)^2},$$

where $\sigma_{t,read}^{2\,adj}$ was the variance of true adjusted school value-added in reading in year *t*, $\sigma_{t,math}^{2\,adj}$ was the variance of true adjusted school value-added in math in year *t*, θ_t^{adj} was the covariance between true school value-added in reading and true school value-added in math, and the other variables had the same values used in (B12a) and (B12b).

Shrinkage weights for the adjusted reading estimates were constructed as

$$\text{(B22a)}\quad \gamma_{st,read}^{adj\,read} = \frac{\left(\sigma_{t,math}^{2\,adj} + V_{st,math}^{error}\right)\sigma_{t,read}^{2\,adj} - \left(\theta_t^{adj} + C_{st}^{error}\right)\theta_t^{adj}}{\left(\sigma_{t,math}^{2\,adj} + V_{st,math}^{error}\right)\left(\sigma_{t,read}^{2\,adj} + V_{st,read}^{error}\right) - \left(\theta_t^{adj} + C_{st}^{error}\right)^2}$$

$$\text{(B22b)}\quad \gamma_{st,math}^{adj\,read} = \frac{\left(\sigma_{t,read}^{2\,adj} + V_{st,read}^{error}\right)\theta_t^{adj} - \sigma_{t,read}^{2\,adj}\left(\theta_t^{adj} + C_{st}^{error}\right)}{\left(\sigma_{t,math}^{2\,adj} + V_{st,math}^{error}\right)\left(\sigma_{t,read}^{2\,adj} + V_{st,read}^{error}\right) - \left(\theta_t^{adj} + C_{st}^{error}\right)^2}$$

Shrunk adjusted school value-added ratings for math and reading were then generated as

$$\text{(B23)}\quad \hat{v}_{spt,math}^{shrunk} = \gamma_{st,math}^{adj\,math}\hat{v}_{spt,math} + \gamma_{st,read}^{adj\,math}\hat{v}_{spt,read}$$

$$\text{(B24)}\quad \hat{v}_{spt,read}^{shrunk} = \gamma_{st,read}^{adj\,read}\hat{v}_{spt,read} + \gamma_{st,math}^{adj\,read}\hat{v}_{spt,math}\,.$$

Years in which the study calculated adjusted school value-added estimates. Adjusted school value-added ratings were calculated only for the years listed in the far right column of table B2. The full measurement period for each principal consisted of a three-year period: a two-year period to estimate a school's current two-year contributions to student growth and the year immediately preceding the two-year period to estimate the baseline school value-added.

Table B2. Years in which adjusted school value-added was estimated

Validation period	Measurement period	Year of adjusted value-added estimate
2007/08–2009/10	2010/11–2012/13	2013
2008/09–2010/11	2011/12–2013/14	2014
2010/11–2012/13	2007/08–2009/10	2010
2011/12–2013/14	2008/09–2010/11	2011

Source: Authors' compilation.

Although the validation analyses required only adjusted school value-added ratings for the years indicated in table B2, the estimation of equation B18 pooled together all available data years (2009/10–2013/14) from which E_{spt} could be obtained. The data were pooled to increase the precision of the estimated coefficients.

School value-added

A measure of school value-added is a common component of principal evaluation systems (examples of states implementing school value-added in principal evaluations include Ohio, Pennsylvania, and Florida). These measures of school value-added typically do not adjust for school value-added from prior years. To assess the validity of school value-added as a measure of principal performance, the study also generated unadjusted school value-added estimates.

To generate unadjusted school value-added ratings, the study followed the same initial approach used to construct adjusted estimates. First, the study estimated school value-added separately by grade and subject for each school-year combination. Then, the study calculated an average school value-added estimate for each subject-school-year combination, where each grade was weighted by the effective number of students in the grade. The study used the same parameters guiding the average adjusted school value-added estimates to determine inclusion of grades in the weighted averages. Finally, the unadjusted school value-added estimates were shrunk using the approach described earlier in this appendix.

Adjusted average achievement

Adjusted average achievement was calculated as the difference between the current average achievement at a principal's school and what it was predicted to be based on average achievement at the same school from a prior year.

To estimate adjusted average achievement, the study followed steps that were analogous to those used to estimate adjusted school value-added. First, mean student achievement was estimated, separately by grade, year, and subject. Next, achievement estimates were averaged across grades, within each subject-year combination. Finally, achievement estimates were adjusted to account for prior average achievement at the school and were transformed into empirical Bayes estimates to account for measurement error. The remainder of this section describes each of these steps in detail.

The study estimated mean student achievement as the weighted average of student assessment scores within each grade, school, year, and subject:

$$\text{(B25)} \quad E_{gstk}^{ach} = \frac{\sum_i A_{igstk} W_{igst}}{\sum_i W_{igst}}.$$

As described earlier, A_{igstk} was the assessment score for student *i* in grade *g* attending school *s* in year *t* for subject *k*, expressed as a *z*-score with mean 0 and standard deviation 1 within each subject-grade-year combination. The achievement estimates used the same student-level weights applied in the value-added estimates.

Because the achievement estimates were measured with error and required adjustment to account for this measurement error, the variances and covariances of the estimation errors for each achievement estimate were calculated. The study applied the same general procedure used to calculate the variance of estimation error for the value-added estimates:

$$\text{(B26)} \quad V_{gstk}^{ach,error} \equiv Var\left(E_{gstk}^{ach}\right) = \left(V_{gtk}^{outcome}\right)\left(Q_{gstk}\right) / \left(W_{gstk}^{2}\right),$$

where $V_{gtk}^{outcome}$ was the average within-school variance of the assessment scores $\left(A_{igstk}\right)$ for grade *g* and subject *k* in year *t*; Q_{gstk} was the sum of squared weights for students in school *s* who had a test score in grade *g*, year *t*, and subject *k*; and W_{gstk} was the total weight for students in school *s* who had a test score in grade *g*, year *t*, and subject *k*.

Likewise, the study estimated the covariance between the estimation errors of each school's math and reading achievement estimates in the same year and grade:

$$\text{(B27)} \quad C_{gst}^{ach,error} \equiv Cov\left(E_{gst,math}^{ach}, E_{gst,read}^{ach}\right) = \frac{\left(C_{gt}^{outcome}\right)\left(Q_{gst}^{both}\right)}{\left(W_{gst,math}\right)\left(W_{gst,read}\right)},$$

where $C_{gt}^{outcome}$ was the average within-school covariance of the math and reading assessment scores for grade *g* in year *t*, and Q_{gst}^{both} was the sum of squared weights for students in school *s* who had both math and reading test scores in grade *g* and year *t*.

The study used the same parameters described in the previous section to determine inclusion of grades in the weighted averages, as calculated in equation B9. Next, the study generated shrunk estimates of average achievement (and estimates of their estimation error variances) using the approach in equations B12a–B17. To calculate adjusted average achievement, the study regressed average achievement on the shrunk estimate of average achievement obtained two years earlier (equation B18). The estimated relationship was then used to adjust for pre-existing conditions at principals' schools (equation B19). The estimated relationships, both for the sample restricted only to principals in their second year at their schools and for the full sample, are shown in table B3.

Finally, the adjusted average achievement ratings were shrunk using the same approach used for the adjusted school value-added estimates in equations B21a–B24.

Table B3. Estimated relationships between average achievement and the same measure obtained two years earlier

Subject	Estimated relationship using all schools	Estimated relationship using schools whose principals were in their second year
Math	0.95* (0.006) [10,387]	0.96* (0.017) [921]
Reading	0.95* (0.006) [10,387]	0.96* (0.015) [921]

* Significant at *p* = 0.05.

Note: Standard errors are indicated in parentheses. Number of school-year combinations are indicated in brackets. Each coefficient estimate is derived from a separate regression in which the dependent variable is a school's unshrunk performance estimate and the independent variable is the school's shrunk performance estimate in the same subject from two years earlier. Every regression also controls for year indicators. Outcome years include 2009/10–2013/14.

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

Average achievement

Several states, such as New York, also incorporate in principal evaluation systems a measure of average achievement that does not adjust for average achievement before a principal's tenure at the school. As such, the study also sought to generate unadjusted average achievement measures to assess their validity as a measure of principal performance.

To generate unadjusted average achievement ratings, the study followed the same initial approach used to construct adjusted estimates. First, the study generated achievement estimates for each unique grade, subject, school, and year combination. An average achievement estimate was calculated for each subject-school-year combination, weighting each grade by the effective number of students in the grade. The study used the same parameters guiding the average estimates for all previously described measures to determine inclusion of grades in the weighted averages. Finally, the unadjusted average achievement ratings were shrunk following the procedure in equations B12a–B17.

Estimating the stable part of each principal's performance rating

Because the principal performance ratings (from the measurement period) were obtained one or more years before or after principals' contributions to student achievement were estimated (in the validation period), the study sought to account for potential drift in measured principal performance over time. Accounting for drift effectively identified the stable part of each principal's performance rating. Analyzing the predictive validity of just the stable part of each measure (see figure 6 of the main report) was a key step toward summarizing the predictive validity of each original, full measure (see figure 7 of the main report).

The study adapted a method Chetty, Friedman, and Rockoff (2014) used to account for drift. First, the study estimated the correlations between annual performance estimates over different time increments, separately for each of the four performance measures. The study then used the estimated associations to generate performance ratings for each principal that adjusted for the instability of the ratings over time and, therefore, isolated the stable part of the principal's annual performance rating.

Estimating the correlations between principal performance ratings obtained in different years. To prepare to examine drift in measured performance over time, the study first removed from the sample all principals in their first year at a school. The performance measures as constructed were intended to reflect a principal's contribution to a school over a two-year period. However, a performance rating for a principal in his or her first year at a school would reflect only one year of the principal's performance at the school, along with one year of the preceding principal's performance. In addition to eliminating principals with only one year of tenure at a school, the dataset also excluded principals who were part of the final predictive validity analysis sample so that no data from those principals' validation periods contributed to any of their performance ratings. Finally, all years of performance ratings for each principal were pooled and the following regression models were estimated, separately for each type of performance measure and for each time interval (indexed by *r* = 1, 2, 3, 4, 5):

(B28a) $E_{pt} = \beta_0 + \beta_{1,(-r)} E_{p(t-r)}^{shrunk} + v_{spt}$

(B28b) $E_{pt} = \beta_0 + \beta_{1,(+r)} E_{p(t+r)}^{shrunk} + v_{spt}$,

where E_{pt} was the unshrunk rating of the performance of principal *p* in year *t*, $E_{p(t-r)}^{shrunk}$ was the shrunk performance rating of the same school from *r* years prior to that year, and $E_{p(t+r)}^{shrunk}$ was the shrunk performance rating of the same school from *r* years after that year. The error term, v_{pt} , captured the transitory component of principal performance over time.

The estimated coefficients from equations B28a and B28b are presented in table B4. Relative to more distant years, performance ratings in years directly preceding or following the current year were better predictors of performance ratings in the current year. The decline in the magnitude of the coefficients as the time before or after the current year increases, for most performance measures, underscored the importance of generating principal ratings that accounted for this drift over time.

Table B4. Estimated relationships between principal performance measures and the same measures obtained in different years

Year of the independent variable	Estimated relationship between a dependent variable consisting of the measure from year *t* and an independent variable consisting of the measure from the row year				Number of principal-year combinations
	Average achievement	Adjusted average achievement	School value-added	Adjusted school value-added	
Math					
Year *t* – 5	0.95*	na	0.25*	na	454–455
Year *t* – 4	0.95*	–0.15*	0.23*	0.15*	576–1,180
Year *t* – 3	0.97*	–0.05*	0.32*	0.20*	1,442–2,195
Year *t* – 2	0.97*	–0.20*	0.44*	0.18*	2,631–3,521
Year *t* – 1	0.98*	0.37*	0.67*	0.46*	4,186–5,272
Year *t* + 1	0.91*	0.37*	0.68*	0.48*	4,186–5,272
Year *t* + 2	0.83*	–0.22*	0.44*	0.17*	2,631–3,521
Year *t* + 3	0.75*	–0.07*	0.32*	0.20*	1,442–2,195
Year *t* + 4	0.72*	–0.17*	0.25*	0.20*	576–1,180
Year *t* + 5	0.68*	na	0.28*	na	454–455
Reading					
Year *t* – 5	0.97*	na	0.26*	na	454–455
Year *t* – 4	0.97*	–0.12*	0.25*	0.18*	576–1,180
Year *t* – 3	0.96*	–0.16*	0.28*	0.16*	1,442–2,195
Year *t* – 2	0.96*	–0.31*	0.40*	0.14*	2,631–3,521
Year *t* – 1	0.98*	0.35*	0.65*	0.46*	4,186–5,272
Year *t* + 1	0.93*	0.36*	0.65*	0.49*	4,186–5,272
Year *t* + 2	0.87*	–0.33*	0.41*	0.16*	2,631–3,521
Year *t* + 3	0.81*	–0.17*	0.29*	0.19*	1,442–2,195
Year *t* + 4	0.78*	–0.15*	0.25*	0.29*	576–1,180
Year *t* + 5	0.78*	na	0.32*	na	454–455

* Significant at $p = 0.05$.

na is not available.

Note: The dependent variable is unshrunk and the independent variable is shrunk. Outcome years include 2008/09–2013/14. The sample excludes both principals in their first year at a school and all principals included in the final analysis sample.

Source: Authors' calculations based on data provided by the Pennsylvania Department of Education, 2006/7–2013/14.

Calculating the stable part of a principal's performance rating. Because principals' performance drifted over time, their single-year performance ratings from the measurement period could not have been the best linear predictors of their expected performance on that measure in the validation period (see figure D1 in appendix D for the combination of measurement and validation periods used in this study). Therefore, those single-year performance ratings could also not have been good predictors of principals' actual contributions to student achievement in the validation period.

To make better predictions in the validation period, the study isolated the part of each principal's performance rating from the measurement period that was expected to persist to the validation period. The following steps were applied to each principal in the final predictive validity analysis. First, the study calculated the interval of time from the year of the performance rating to the year of the validation period in which the performance rating was supposed to predict principals' contributions. Next, if the measurement period year occurred *r* years before the validation period year, the stable part of the performance rating was estimated as

(B29a) $E_{pt}^{drift-r} = \hat{\beta}_{1,(-r)} E_{pt}^{shrunk}$,

where E_{pt}^{shrunk} was the shrunk performance rating of principal *p* in year t and $\hat{\beta}_{1,(-r)}$ was obtained from equation B28a. Likewise, if the measurement period year occurred *r* years after the validation period, the stable part was estimated as

(B29b) $E_{pt}^{drift+r} = \hat{\beta}_{1,(+r)} E_{pt}^{shrunk}$,

where $\hat{\beta}_{1,(+r)}$ was obtained from B28b.

Due to data constraints, there was never more than a four-year gap in performance ratings that adjusted for prior school performance for any principal. Because the drift coefficients appeared to stabilize at spans longer than two years for all measures and were similar for both the four-year and five-year spans for unadjusted measures, for estimated performance using adjusted measures, the study used the estimated coefficients $\hat{\beta}_{1,(+4)}$ and $\hat{\beta}_{1,(-4)}$ to generate $E_{pt}^{drift+5}$ and $E_{pt}^{drift-5}$, respectively.

Appendix C. Technical details of methods for examining the stability of principal performance ratings

This appendix explains how the study used regression methods to identify the fraction of the variation in single-year performance ratings that was stable over time. It also provides details on the formulas used for simulating the stability of multiyear averages of performance ratings. This study examined the stability of principals' performance ratings regardless of whether they remained at the same schools across years. All measures in the study were intended to ascertain the effectiveness of principals—not schools—so the ratings were supposed to gauge how well principals would perform at whatever schools they were assigned.

Estimating the fraction of the variance in single-year performance ratings that was stable

To estimate the fraction of the variance in single-year performance ratings that was stable, the study began with a general analytic framework that assumed each performance measure had both a stable and transitory component. Formally, a performance rating, E_{pt}, for principal *p* in year *t* was assumed to have a component $u_p^{(d)}$ that was stable for at least a specified duration *d* of interest (such as *d*=4 years) and a component $e_{pt}^{(d)}$ that dissipated before *d* years had elapsed (meaning that $e_{pt}^{(d)}$ was uncorrelated with $e_{p,t+r}^{(d)}$ for $r \geq d$):

(C1) $$E_{pt} = u_p^{(d)} + e_{pt}^{(d)}.$$

The objective was to estimate the fraction $\left(F_{(d)}\right)$ of the variance of E_{pt} attributable to $u_p^{(d)}$ rather than $e_{pt}^{(d)}$:

(C2) $$F_{(d)} \equiv \frac{Var\left(u_p^{(d)}\right)}{Var\left(E_{pt}\right)}.$$

To estimate $F_{(d)}$, the study estimated a linear regression model in which the independent variable was a single-year performance rating $\left(E_{pt}\right)$ and the dependent variable was the same principal's performance rating *d* years later $\left(E_{p,t+d}\right)$:

(C3) $$E_{p,t+d} = \beta_0 + \beta_1 E_{pt} + \varepsilon_{pt}.$$

The coefficient β_1 from the regression model was

(C4) $$\beta_1 = \frac{Cov\left(E_{p,t+d}, E_{pt}\right)}{Var\left(E_{pt}\right)} = \frac{Cov\left(u_p^{(d)} + e_{p,t+d}^{(d)}, u_p^{(d)} + e_{pt}^{(d)}\right)}{Var\left(E_{pt}\right)} = \frac{Var\left(u_p^{(d)}\right)}{Var\left(E_{pd}\right)} = F_{(d)}.$$

In other words, the coefficient from regressing subsequent performance ratings on performance ratings from d years earlier identified the fraction of the variance of the performance ratings that was stable for at least d years.

The report documented stability for $d = 1, 2, 3$, and 4 years (see figures 1 and 2 of the main report). To implement the regression in equation C3 separately for each duration d, the study used shrunk performance ratings as the independent variable and unshrunk performance ratings from d years later as the dependent variable. The resulting estimate of $F_{(d)}$ could be interpreted as the fraction of the variance in principals' *expected* performance on the measure (that is, performance that would be observed if there was no random estimation error attributable to small student samples) that was stable for at least d years. The regression in equation C3 was the same one used in correcting performance ratings for drift, as described in appendix B (equation B28a). Therefore, the results from estimating this regression were already reported in table B4. For example, when using equation C3 to examine stability for 4 years, the results are reported in the row of table B4 labeled (t – 4).

The main report compared the values of $F_{(d)}$ from this study with values found previously by Chetty et al. (2014) for teacher value-added ratings. However, in the values of $F_{(d)}$ reported by Chetty et al. (2014) (see the autocorrelations reported in their table 2), instability could stem from both changes in teachers' true performance across years and fluctuations attributable to limited numbers of students who contribute to the ratings. In the values of $F_{(d)}$ from this study, instability from limited numbers of students has already been removed because the study used shrunk performance ratings as the independent variable in equation C3. Therefore, to obtain comparable measures of the stability of teacher value-added ratings from Chetty et al. (2014), the study team divided their autocorrelations by the reliability of the value-added ratings, where reliability was the proportion of the variance in the value-added ratings not attributable to sampling error.

Simulating the fraction of the variance in multiyear performance ratings that would be stable

Instead of using a single-year performance rating, suppose that a state or district evaluated each principal based on a rolling average of three years of performance ratings. This section derives the formula for simulating the fraction of the variation in the three-year rolling average that would persist to the following year.

Formally, suppose that in any year t a principal is evaluated based on a rolling average, E_{pt}^{roll} , of his or her ratings from years (t – 2), (t – 1), and t:

$$\text{(C5)} \qquad E_{pt}^{roll} = \left(E_{p,t-2} + E_{p,t-1} + E_{pt} \right) / 3 .$$

The objective is to determine the fraction of the variance of E_{pt}^{roll} that would persist to year (*t* + 1), denoted as $F_{(1)}^{roll}$. Based on the discussion in the previous section of this appendix, this fraction would be equivalent to the coefficient from a regression in which E_{pt}^{roll} is the independent variable and the single-year rating from year (*t*+1), $E_{p,t+1}$, is the dependent variable.

However, it is not necessary to estimate an actual regression of $E_{p,t+1}$ on E_{pt}^{roll} to obtain $F_{(1)}^{roll}$. As shown next, the coefficient from such a regression would depend on inputs that will have already been estimated. This is because E_{pt}^{roll} is simply an average of $E_{p,t-2}$, $E_{p,t-1}$, and E_{pt}. Therefore, the relationship between E_{pt}^{roll} and $E_{p,t+1}$ can be derived from the relationship between $E_{p,t-2}$ and $E_{p,t+1}$ (stability for three years), the relationship between $E_{p,t-1}$ and $E_{p,t+1}$ (stability for two years), and the relationship between E_{pt} and $E_{p,t+1}$ (stability for one year)—all of which will have been estimated previously. Simulating $F_{(1)}^{roll}$ based on previous results—rather than regressing $E_{p,t+1}$ on E_{pt}^{roll} directly—is advantageous because it does not require limiting the analysis to principals who have ratings from all of the years being considered.

Formally, consider a hypothetical regression of $E_{p,t+1}$ on E_{pt}^{roll}. Additionally, assume that the single-year performance measure has the same variance in all years—that is, $Var\left(E_{p,t-2}\right)=Var\left(E_{p,t-1}\right)=Var\left(E_{pt}\right)=Var\left(E_{p,t+1}\right)$. The coefficient from the regression would be

$$\text{(C6)}\qquad F_{(1)}^{roll}=\frac{Cov\left(E_{p,t+1},E_{pt}^{roll}\right)}{Var\left(E_{pt}^{roll}\right)}=\frac{Cov\left(E_{p,t+1},\left(E_{p,t-2}+E_{p,t-1}+E_{pt}\right)/3\right)}{Var\left(E_{pt}^{roll}\right)}$$

$$=\frac{\left(\frac{1}{3}\right)Cov\left(E_{p,t+1},E_{p,t-2}+E_{p,t-1}+E_{pt}\right)}{Var\left(E_{pt}^{roll}\right)}=\frac{\left(\frac{1}{3}\right)Cov\left(E_{p,t+1},E_{p,t-2}+E_{p,t-1}+E_{pt}\right)}{Var\left(E_{pt}\right)}\times\frac{Var\left(E_{pt}\right)}{Var\left(E_{pt}^{roll}\right)}$$

$$=\left(\frac{1}{3}\right)\left[\frac{Cov\left(E_{p,t+1},E_{p,t-2}\right)}{Var\left(E_{pt}\right)}+\frac{Cov\left(E_{p,t+1},E_{p,t-1}\right)}{Var\left(E_{pt}\right)}+\frac{Cov\left(E_{p,t+1},E_{pt}\right)}{Var\left(E_{pt}\right)}\right]\times\frac{Var\left(E_{pt}\right)}{Var\left(E_{pt}^{roll}\right)}$$

$$=\left(\frac{1}{3}\right)\left[F_{(3)}+F_{(2)}+F_{(1)}\right]\times\frac{Var\left(E_{pt}\right)}{Var\left(E_{pt}^{roll}\right)}$$

$$=\left(\frac{1}{3}\right)\left[F_{(3)}+F_{(2)}+F_{(1)}\right]\times\frac{Var\left(E_{pt}\right)}{Var\left(\left(E_{p,t-2}+E_{p,t-1}+E_{pt}\right)/3\right)}$$

$$=\left(\frac{1}{3}\right)\left[F_{(3)}+F_{(2)}+F_{(1)}\right]\times$$

$$\frac{9Var\left(E_{pt}\right)}{\left[Var\left(E_{p,t-2}\right)+Var\left(E_{p,t-1}\right)+Var\left(E_{pt}\right)+2Cov\left(E_{p,t-2},E_{p,t-1}\right)+2Cov\left(E_{p,t-2},E_{pt}\right)+2Cov\left(E_{p,t-1},E_{pt}\right)\right]}$$

$$=\left(\frac{1}{3}\right)\left[F_{(3)}+F_{(2)}+F_{(1)}\right]\times$$

$$\frac{9}{\left[\frac{Var\left(E_{p,t-2}\right)}{Var\left(E_{pt}\right)}+\frac{Var\left(E_{p,t-1}\right)}{Var\left(E_{pt}\right)}+\frac{Var\left(E_{pt}\right)}{Var\left(E_{pt}\right)}+\frac{2Cov\left(E_{p,t-2},E_{p,t-1}\right)}{Var\left(E_{pt}\right)}+\frac{2Cov\left(E_{p,t-2},E_{pt}\right)}{Var\left(E_{pt}\right)}+\frac{2Cov\left(E_{p,t-1},E_{pt}\right)}{Var\left(E_{pt}\right)}\right]}$$

$$=\left(\frac{1}{3}\right)\left[F_{(3)}+F_{(2)}+F_{(1)}\right]\times\frac{9}{\left[1+1+1+2F_{(1)}+2F_{(2)}+2F_{(1)}\right]}$$

$$=\frac{3\left(F_{(3)}+F_{(2)}+F_{(1)}\right)}{\left[3+4F_{(1)}+2F_{(2)}\right]}.$$

Equation C6 shows that the fraction of the variance in a three-year rolling average that persists to the following year can be calculated directly from the fraction of the variance in a single-year rating that persists for one, two, and three years.

Appendix D. Technical methods for assessing the predictive validity of principal performance measures

This appendix provides technical details of the study's approach to assessing the predictive validity of principal performance measures. As mentioned in box 2 of the main report, the study examined the predictive validity of each performance measure in two stages. First, the study examined the extent to which the stable part of each performance measure was associated with principals' contributions to student achievement in a different time period. This initial analysis focused on the stable part of a measure because only stable ratings had any potential to carry information about principals' contributions in other years. As discussed later in this appendix, the amount of time that elapsed between a principal's initial rating and the final assessment of a principal's contributions depended on the available data and, therefore, varied across principals.

The second stage of the analysis adapted the empirical findings from the first stage to summarize the predictive validity of each measure in a way that would be most relevant to personnel decisions. Specifically, the study simulated the extent to which each original, full measure could predict principals' contributions in the following year. This analysis focused on the original measures—encompassing all ratings, not just the stable ones—assuming that an evaluation system would not use just the stable ratings, which cannot reflect real evolution in principals' performance. This analysis also assumed that state and district officials would be most interested in predicting principals' contributions in the following year when making end-of-year personnel decisions for the subsequent school year.

The first stage of the analysis—assessing the predictive validity of the stable part of each measure—entailed carrying out three steps:

- **Step 1: Calculating performance ratings based on each measure.** The study implemented each of four performance measures to calculate performance ratings for principals in a specified period, referred to as the measurement period. As described earlier in the report, the four performance measures examined in the study were those that a state or district could realistically and broadly apply to evaluate most of its principals. A key objective of the study was to ascertain how well these performance measures predicted principals' contributions to student achievement.

- **Step 2: Obtaining benchmark estimates of principals' contributions to student achievement.** The study used a quasi-experimental method—distinct from the four performance measures examined in the study—to obtain estimates of principals' contributions to student achievement in an entirely different period, referred to as the validation period. The estimates in the validation period needed to represent the most rigorous available estimates of principals' contributions so that the four performance measures could be judged by how well they predicted the benchmark estimates. For this reason, the study referred to the estimates from the validation period as "benchmark" estimates of principals' contributions to student achievement.

- **Step 3: Assessing the relationship between the stable part of each performance measure and the benchmark estimates of principals' contributions to student achievement.** The study assessed the relationship between principals' stable performance ratings from the measurement period and the benchmark estimates of principals' contributions from the validation period. The study referred to the magnitude of this relationship as the predictive validity of the stable part of each performance measure.

A critical feature of the study design was to generate benchmark estimates from an entirely separate time period as the period used to generate performance ratings from the four measures. This reduced the risk that transient influences on schools that were outside of principals' control would falsely inflate the relationship between the performance ratings and the benchmark estimates. If, instead, performance ratings and benchmark estimates were based on the same years of data, then transient influences that increased or decreased principals' performance ratings would, at the same time, increase or decrease the benchmark estimates. For example, if a brutal winter led to an unusually large number of snow days and decreased schools' test scores in a specific year, the principal's performance rating on any of the four measures and the benchmark estimate of the principal's contributions would be lower than warranted in that year. Performance ratings would then seem to predict the benchmark estimates for reasons other than the quality of the performance measure. The study design removed this type of bias from the analysis of predictive validity.

The second stage of the analysis extended the first stage with one additional step:

- **Step 4: Summarizing the extent to which each original, full performance measure could predict principals' contributions to student achievement in the following year.** A performance measure could more accurately predict principals' contributions in the following year to the extent that the measure was more stable between consecutive years (from the analyses described in appendix C) and the stable part of the measure was more strongly associated with principals' contributions (from Step 3 above).

The specific performance measures that the study implemented in Step 1 were described earlier in appendix B. Steps 2 through 4, which were similar regardless of the performance measure being considered, are the focus of this appendix. The remainder of this appendix describes the methods for obtaining benchmark estimates of principals' contributions to student achievement in the validation period, the approach to measuring the relationship between stable performance ratings and benchmark estimates, the specific time periods that the study used, the number of principals included in the analysis, and the formulas for summarizing a performance measure's accuracy for predicting principals' contributions in the following year.

Obtaining benchmark estimates of principals' contributions to student achievement in the validation period

To serve as an appropriate benchmark for assessing the predictive validity of performance measures, estimates of principals' contributions to student achievement in the validation period need to be unbiased. In the ideal scenario, these estimates would come from a random assignment design. For example, two studies of the predictive validity of teacher value-added measures randomly assigned teachers to classrooms to obtain unbiased estimates of teachers' contributions to student achievement in the studies' validation periods (Kane, McCaffrey, Miller, & Staiger, 2013; Kane & Staiger, 2008). However, random assignment of principals to schools would have been politically and operationally infeasible on the scale needed for this study.

Instead, this study used a quasi-experimental approach to estimate differences between principals in their contributions to student achievement. This approach used leadership transitions—instances in which a principal replaced another at the same school—as the basis for comparing the successor's and predecessor's contributions to student achievement. In what follows, this section describes the general rationale for this approach and then lays out its technical details.

Leadership transitions as the basis for revealing principals' effectiveness. Turnover of personnel within an organization can shed light on differences between the successors' and predecessors' influence on that organization. In their seminal study of the predictive validity of teacher value-added measures, Chetty, Friedman, and Rockoff (2014) estimated how test scores changed within instructional teams when incoming teachers replaced departing ones. The authors regarded these test-score changes as unbiased estimates of the relative effectiveness of the incoming and departing teachers, under the assumption that teachers' arrival and departure were unrelated to other factors that might have increased or decreased test scores in these instructional teams. They found that these test-score changes could be accurately predicted by the teachers' relative value-added scores from other years.

This study adapted the approach of Chetty et al. (2014) to estimate how departing principals differed from their successors in their contributions to student achievement. A comparison of the school's test scores before and after the arrival of the successor estimated the successor's contribution relative to that of the predecessor.

A major advantage of comparing principals who served at the *same* school is that it holds constant many factors outside of principals' control that differ *between* schools. Schools differ from each other on several factors that might influence test scores, such as neighborhood safety, social cohesion among the students' families, the families' socioeconomic circumstances, commuting times for teachers, and the quality of school facilities. Under the assumption that these factors are persistent over time, they cannot be responsible for test-score changes during a leadership transition, so the study could attribute these changes to differences between the predecessor's and successor's contributions to student achievement.

Nevertheless, if other factors at a school change in a way that is systematically related to whether a principal is replaced by a better or worse successor, then test-score changes might reflect these confounding factors—not just the principals' relative effectiveness. For example, if districts purposely replace incumbent principals with better ones at the same time that they add other resources or interventions to the school, then differences in the principals' contributions might be falsely obscured or magnified. The study's approach to checking for and addressing these potential threats is addressed in appendix F.

Many studies on principal effectiveness have used leadership transitions to estimate differences between principals in their contributions to student achievement (Branch, Hanushek, & Rivkin, 2012; Cannon, Figlio, & Sass, 2012; Chiang, Lipscomb, & Gill, 2016; Coelli & Green, 2012; Dhuey & Smith 2013, 2014; Grissom, Kalogrides, & Loeb, 2015; Lipscomb, Chiang, & Gill, 2012). Researchers have preferred this approach because, as discussed earlier, it controls for the persistent attributes of schools that shape student outcomes. However, this method cannot be the basis for a real evaluation system because it can be applied only to principals involved in leadership transitions, and it compares each principal to only one other principal rather than to a broad reference group. Therefore, this method was not one of the performance measures that the study implemented in the measurement period, as those measures were intended to be candidates for use in real evaluation systems. Nevertheless, for the subset of principals involved in leadership transitions in the validation period, the method provided the most rigorous available benchmark with which to judge the predictive validity of performance estimates from the measurement period.

Benchmark estimator for the difference in principals' contributions to student achievement. To estimate the contribution of an incoming principal at a school relative to that of a departing principal, the study calculated the school's average test score in the incoming principal's second year and subtracted the school's average test score in the departing principal's final year. Because test scores were standardized into *z*-scores, relative contributions were expressed in standard deviations of student test scores. The study constructed separate estimates for reading and math.

A key element of this approach was to measure a school's test scores under an incoming principal after he or she has led the school for two years. Prior research has shown that a principal's full impact on a school's performance materializes over several years (Coelli & Green, 2012). For example, effective principals could improve their schools' test scores by encouraging low-performing teachers to leave and filling those vacancies with higher-performing teachers, but this process might take time. Incoming principals might not be able to make any staffing changes at all in their first year, especially if they are hired soon before the start of the school year. A principal's second year at a school is likely to be the earliest year in which the school's test scores could reflect at least some staffing changes resulting from his or her leadership.

Waiting three or more years before measuring a school's performance under a new principal would potentially capture even more of the principal's influence on the school. However, the available data for this study were insufficient to permit a waiting period of three or more years. As discussed in appendix A, data that linked students to schools spanned seven years. If the study required three years to elapse under an incoming principal's leadership in the validation period, then any validation period must span four total years (one year under the departing principal and three years under the incoming principal. A measurement period must span the same number of years as a validation period so that both periods estimate principals' contributions over the same duration of leadership. Because the two periods cannot overlap, the seven-year dataset could not accommodate a four-year measurement period plus a four-year validation period.

Nevertheless, prior evidence suggests that two years are sufficient for incoming principals to distinguish themselves from their predecessors reasonably clearly, even if not fully. Using data from British Columbia, Coelli and Green (2012) estimated that after two years, the difference between a school's test scores under a new principal versus his or her predecessor was 40 to 50 percent of the full, long-run difference.

Therefore, the estimated difference between an incoming and departing principal's contributions to student achievement must be interpreted properly. It estimates the difference in students' test scores due to the incoming principal's first two years of leadership at the school compared with the scores they would have achieved if the departing principal had stayed for two more years. It does not reflect the scenario in which the students' exposure to the incoming principal had occurred when this principal had already served at the school a long time.

Grades included in the estimates. Average test scores in both the departing principal's last year and the incoming principal's second year were based on students enrolled in grades 5 to 8 in those years. Although annual testing in Pennsylvania began in grade 3, grade 5 is the earliest grade at which students who were tested in the incoming principal's second year would have baseline (pre-transition) scores from two years earlier. As discussed later in appendix F, comparing the pre-transition scores of students who were and were not exposed to the incoming principal was a key check for potential bias in the estimates.

In addition, the study excluded a school's lowest grade if it fell into the grade 5–8 range. By the end of the incoming principal's second year, students in the school's lowest grade had experienced only one year of this principal's leadership, whereas students in all other grades were expected to have experienced two years. Excluding the school's lowest grade resulted in a consistent level of students' expected exposure (of two years) to the incoming principal.[9]

[9] For some students, actual exposure to the incoming principal could be less than the expected exposure based on their grade level. Students who moved into a school after the incoming principal began leading the school experienced fewer than two years of the principal's leadership. Nevertheless, these students were included in the estimates, because excluding them would have resulted in estimates that pertained only to non-movers, who were not representative of the full population of students in

Method of averaging. To calculate average test scores at a school in both the departing principal's last year and the incoming principal's second year, the study calculated the average in two steps. First, the study calculated average test scores separately in each grade level of the school, with students weighted by the fraction of the school year in which they were enrolled in the school. Second, the study then took an average across grade levels at the school, with grades weighted by the effective number of students (that is, the total student weight in the grade) from the departing principal's final year. Therefore, grades were weighted in an identical manner for the departing and incoming principal, which prevented differences in average scores from being driven artificially by differences in the grade distribution at the school.

Assessing the relationship between stable performance ratings from each measure and the benchmark estimates

After estimating differences in principals' contributions to student achievement in the validation period, the ultimate objective was to assess how well they could be predicted by differences in performance ratings from the measurement period. Among the pairs of principals involved in leadership transitions in the validation period, the study identified those in which both principals in the pair also had performance ratings from an entirely separate measurement period while they were leading different schools. Using those pairs of principals, the study estimated a regression in which the within-pair difference in the stable part of their performance ratings (from the measurement period) was the independent variable for predicting the within-pair difference in contributions to student achievement (based on benchmark estimates from the validation period).

Formally, let $\hat{Y}_{pk}$ be the average achievement of students tested under principal *p* in the validation period in subject *k* (either in a departing principal's last year or an incoming principal's second year). In addition, let $\hat{M}_{pk}$ be the stable part of the principal's performance rating from the measurement period. Using a dataset that stacked reading and math observations, the study estimated the following regression:

$$\text{(D1)} \qquad \hat{Y}_{pk} = \beta_0 + \beta_1 \hat{M}_{pk} + \beta_2 Math_k + \beta_3 I_p + \beta_4 \left(Math_k \times I_p \right) + D_{pk}\gamma + X_{pk}\delta + \varepsilon_{pk} \,,$$

where $Math_k$ was a dichotomous indicator for math observations, I_p was a dichotomous indicator for incoming principals, D_{pk} was a vector of pair-by-subject indicators, X_{pk} was a vector of covariates measuring the average background characteristics of students tested under the principal; ε_{pk} was a random error term, which included measurement error associated with $\hat{Y}_{pk}$, and β_0, β_1, β_2, β_3, β_4, γ, and δ were parameters to be estimated.

the study schools.

In equation D1, the coefficient of interest, β_1, represented the extent to which principals' stable performance ratings on each of the four measures predicted the benchmark estimates of their contributions to student achievement. This study refers to β_1 as the prediction coefficient. The key independent variable of equation D1, $\hat{M}_{pk}$, was expressed in the same units as the dependent variable—standard deviations of student test scores. Therefore, a prediction coefficient of one would represent ideal predictive validity—the scenario in which the stable difference in performance ratings between two principals fully predicted their contributions to student achievement. More generally, the prediction coefficient answered the following question: What fraction of the stable difference in performance ratings between two principals reflected the difference in their contributions to student achievement in another year? Because deviations of the prediction coefficient from one reflect bias in the performance estimates, Chetty et al. (2014) refer to $\left(1-\beta_1\right)$ as forecast bias.

Including the indicator for incoming principals, I_p, in equation D1 accounted for the possibility that test scores after a transition would systematically differ from those before a transition for reasons outside of principals' control. For example, prior research has found that a principal transition typically coincides with a dip in student achievement at a school that begins before a transition and continues for about two years after a transition (Miller, 2013). This suggests that student achievement under an incoming principal may consistently be lower than under an outgoing principal for reasons unrelated to the principals' long-run effectiveness. Inclusion of I_p accounts for these types of unobserved differences between pre-transition and post-transition years that affect all schools in the study. Appendix F also presents supplemental analyses that account for systematic differences between pre-transition and post-transition years that depend on the pre-transition level of achievement at a school.

The covariates, X_{pk}, were included in equation D1 to improve precision. These covariates accounted for random variation in a school's average test scores due to random fluctuations in the background characteristics of the tested students.[10] Each covariate pertained to a specified background characteristic and was structured as the average value of the characteristic for students tested at the school under the principal. The background characteristics consisted of binary indicators for male gender, black race, Hispanic ethnicity, free or reduced-price meals recipients, English learner students, special education students, mobile students (students who changed schools during the school year), and students who were overage for their grade.

Each principal had at least two observations in the dataset—one for math and one for reading. In addition, some principals belonged to multiple pairs, serving as the departing principal

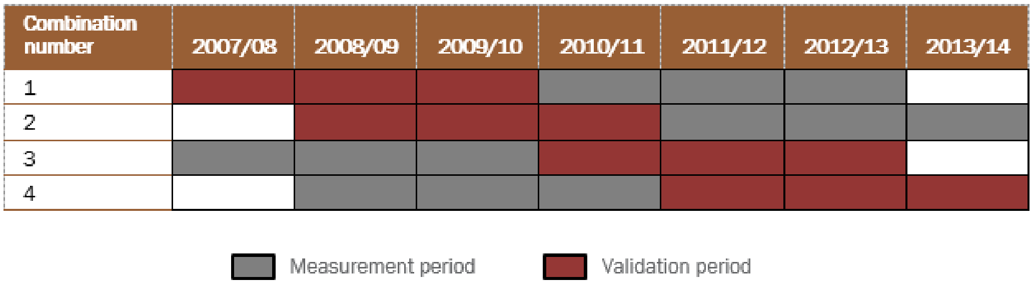
[10] An alternative (and conceptually equivalent) approach would have been to control for these fluctuations in student background characteristics when constructing the benchmark estimates, $\Delta\hat{E}_{pk}$. This study followed the approach of Chetty et al. (2014) in making the benchmark estimates as simple and transparent as possible.

for one pair and the incoming principal for another pair. In estimation of equation D1, robust standard errors accounted for the clustering of multiple observations for the same principal.

Time periods included in the analysis

Each pair of principals in the study had performance ratings from a measurement period and benchmark estimates of their contributions to student achievement in a validation period. Combinations of measurement and validation periods varied across pairs of principals. As discussed earlier, each validation period encompassed a leadership transition and spanned three years—the final year of the departing principal and the first two years of the incoming principal. Each measurement period also needed to span three years so that the two periods assessed principals' contributions to student achievement over the same duration of leadership. In total, the final sample included four distinct combinations of measurement and validation periods (figure D1).

Figure D1. Combinations of measurement and validation periods the study examined



Combination number	2007/08	2008/09	2009/10	2010/11	2011/12	2012/13	2013/14
1	🟥	🟥	🟥	▨	▨	▨	
2		🟥	🟥	🟥	▨	▨	▨
3	▨	▨	▨	🟥	🟥	🟥	
4		▨	▨	▨	🟥	🟥	🟥

▨ Measurement period 🟥 Validation period

Source: Authors' calculations based on job assignment data provided by the Pennsylvania Department of Education, 2007/8–2013/14.

An example of a principal pair whose measurement and validation periods came from the final row in figure D1 is presented in figure D2. Principals 1 and 2 led separate schools—denoted by schools A and B, respectively—in the final two years of their measurement period, 2009/10 and 2010/11, and received a single-year performance rating from 2010/11 on each measure. (The first year of the measurement period, 2008/09, was used only to measure the baseline performance of schools A and B; see appendix B.) The first year of the validation period, 2011/12, was also principal 1's final year at school A. Principal 2 then transferred to school A and began leading it in 2012/13. In principal 2's second year at school A (2013/14), the study measured average student achievement and compared it to the average student achievement in principal 1's final year at that school (2011/12), producing the benchmark estimate of the two principals' relative contributions to student achievement.

Figure D2. Example of a principal comparison in the study

Principal	Measurement period	Validation period
Principal 1 ("predecessor")	Led school A in 2009/10 and 2010/11	Continued to lead school A in 2011/12, then left
Principal 2 ("successor")	Led school B in 2009/10 and 2010/11	Continued to lead school B in 2011/12; began leading school A in 2012/13; remained at school A in 2013/14

Source: Authors' example.

Summarizing each original measure's accuracy for predicting principals' contributions in the following year

The final step of the analysis was to summarize the extent to which each original, full performance measure could predict principals' contributions to student achievement in the following year. This step was based on a simulation that combined findings on the predictive validity of the stable part of each measure (from equation D1) with findings on each measure's stability between consecutive years (from appendix C).

Formally, let δ be the prediction coefficient of each original performance measure for forecasting principals' contributions to student achievement in the following year. In other words, δ answers the following question: What fraction of the original difference in performance ratings between two principals reflects the difference in their contributions to student achievement in the following year? For each measure, the study calculated the value of δ to be

$$\text{(D2)} \qquad \delta = \beta_1 \times F_{(1)},$$

where β_1 was the prediction coefficient of the stable part of the measure (as defined in equation D1 and reported in figure 6 of the main report) and $F_{(1)}$ was the fraction of the variation in performance ratings that persisted between consecutive years (as defined in appendix C and reported in figures 1 and 2 of the main report). For the main analyses that pooled together data on math and reading to estimate β_1, $F_{(1)}$ was also averaged across the two subjects. For supplemental analyses that produced subject-specific estimates of β_1, the study used the value of $F_{(1)}$ in the same subject.

The study calculated the standard error of $\hat{\delta}$ as the standard error of $\hat{\beta}_1$ multiplied by $F_{(1)}$. This approach assumed that the value of $F_{(1)}$ had no statistical uncertainty, which was approximately true given the large sample size used to estimate $F_{(1)}$ (see table B4).

The remainder of this section provides a derivation of equation D2 based on a simple analytic framework. Let E_{pt} denote the original performance rating earned by principal *p* in year *t*. As discussed in appendix C, E_{pt} can be expressed as

$$\text{(D3)} \qquad E_{pt} = u_p^{(1)} + e_{pt}^{(1)} ,$$

where $u_p^{(1)}$ is a component that persists for at least one year, and $e_{pt}^{(1)}$ is a component that is uncorrelated across years.

Suppose, hypothetically, that every principal's contribution to student achievement in year (*t*+1), denoted by $C_{p,t+1}$, could be observed. In this scenario, each measure's prediction coefficient (δ) for forecasting principals' contributions in the following year could be obtained from an ordinary least squares regression of $C_{p,t+1}$ on E_{pt}. In expectation, this coefficient would be

$$\text{(D4)} \qquad \delta = \frac{Cov\left(C_{p,t+1}, E_{pt}\right)}{Var\left(E_{pt}\right)} = \frac{Cov\left(C_{p,t+1}, u_p^{(1)} + e_{pt}^{(1)}\right)}{Var\left(E_{pt}\right)} = \frac{Cov\left(C_{p,t+1}, u_p^{(1)}\right)}{Var\left(E_{pt}\right)}$$

$$= \frac{Cov\left(C_{p,t+1}, u_p^{(1)}\right)}{Var\left(u_p^{(1)}\right)} \times \frac{Var\left(u_p^{(1)}\right)}{Var\left(E_{pt}\right)} = \beta_1 \times F_{(1)} ,$$

as claimed in equation D2.

A version of equation D2 was also used to assess the predictive validity of a three-year average of performance ratings for forecasting principals' contributions in the following year. However, in this case, $F_{(1)}$ was replaced by $F_{(1)}^{roll}$, the fraction of the variation in the three-year average that persisted to the following year.

Statistical precision of the analysis

The final study sample consisted of 123 distinct principals, 134 principal-year combinations (grouped into 67 pairs of incoming and departing principals), and 268 principal-year-subject combinations. Although this sample size yielded poor levels of precision for estimating prediction coefficients of the stable parts of the measures, it yielded moderate levels of precision for estimating the extent to which the original, full measures (either from a single year or three-year rolling average) could predict principals' contributions in the following year.

A useful measure of the statistical uncertainty in the estimated prediction coefficient is the margin of error, defined as half of the width of the 95 percent confidence interval. The margin of error for the prediction coefficients reported in figures 6 through 8 of the main report are shown in table D1. The study had high margins of error for estimating prediction coefficients of the stable parts of school value-added (0.44) and adjusted school value-added (0.79). However, the stable parts represented only a portion of the original measures, with the remaining (unstable) portion assumed to have zero predictive validity with complete certainty. Therefore, precision was better when estimating prediction coefficients of the original measures to forecast principals' contributions in the following year; for the value-added measures, margins of error ranged from 0.28 to 0.36.

As a point of comparison, this study had noticeably worse precision than that of Chetty et al. (2014), whose analysis had a margin of error of 0.07. On the other hand, this study's precision for assessing the predictive validity of the original measures to forecast principals' contributions in the following year was similar to that of Kane et al. (2013), whose analysis had a margin of error of 0.25.

Table D1. Expected margin of error in estimates of prediction coefficients

	Margin of error for the prediction coefficient of the		
Principal performance measure	Stable part of the measure	Full measure, when predicting the next year's contribution	Three-year rolling average of the full measure, when predicting the next year's contribution
Average achievement	0.14	0.14	0.14
Adjusted average achievement[a]			
School value-added	0.44	0.29	0.28
Adjusted school value-added	0.79	0.36	0.36

Note: The margin of error is defined as half of the width of the 95 percent confidence interval.

[a] The adjusted average achievement measure had no stable part.

Source: Authors' calculations based on data provided by the Pennsylvania Department of Education, 2006/7—2013/14.

Appendix E. Study sample for assessing predictive validity

This appendix provides details on how the study team constructed the final analysis sample of principals used for assessing the predictive validity of principal performance measures. It also describes the characteristics of those principals and the students in their schools.

Analysis sample of principals

The predictive validity analysis included 123 Pennsylvania principals for whom it was possible to examine the extent to which differences in their performance ratings predicted differences in their contributions to student achievement, using the approach described in appendix D. These principals were either the outgoing or incoming principals during school leadership changes that occurred in specific years, with both the predecessor and the successor also needing to be leading a school in a specific set of different years.

The beginning sample for the study included all Pennsylvania principals from 2007/08–2013/14 whose school had at least one grade in the 5 to 8 range. From this group, the study team included only principals meeting the following two conditions during the school years in which they served:

- Condition 1: The principal was the only leader of his or her school during a school year.
- Condition 2: The principal led only one school during a school year.

The study team further restricted the sample to retain only those principals for whom it was possible to calculate their contribution to student achievement. These principals needed to be involved in a school leadership transition during specific school years. In particular, the study retained:

- Condition 3: Principals who began leading a school during 2008/09, 2009/10, 2011/12, or 2012/13; who led the same school the following year; and whose predecessor was also in the sample the first two conditions.
- Condition 4: The predecessors of the principals described in Condition 3.

The study design required that the principals in Conditions 3 and 4 also be leading schools in specific other years, during which their performance estimates would be obtained. The final sample used for the prediction analysis included the set of these principals who met one of the requirements below:

- *For the principals who began leading a school in 2008/09 and their predecessor who left in* 2007/08, both in the pair also led a school serving at least one grade in the 5 to 8 range in 2011/12 and 2012/13.
- *For the principals who began leading a school in 2009/10 and their predecessor who left in* 2008/09, both in the pair also led a school serving at least one grade in the 5 to 8 range in 2012/13 and 2013/14.

- *For the principals who began leading a school in 2011/12 and their predecessor who left in 2010/11*, both in the pair also led a school serving at least one grade in the 5 to 8 range in 2008/09 and 2009/10.
- *For the principals who began leading a school in 2012/13 and their predecessor who left in 2011/12*, both in the pair also led a school serving at least one grade in the 5 to 8 range in 2009/10 and 2010/11.

How these requirements affected the size of the principal sample that could be included in the predictive validity analysis is shown in table E1. The analysis included 134 principal-year combinations each for math and for reading (268 principal-year-subject observations in the predictive validity analysis sample). These 134 represented 123 distinct principals, as some principals served as both an incoming and a departing principal across separate pairs.

Table E1. How the study constructed the principal sample for the predictive validity analysis

Type of sample restriction	Reduction in sample size	Remaining sample size
Beginning study sample		
All Pennsylvania principals from 2007/08–2013/14 whose school had at least one grade in the 5 to 8 range		4,269
Initial sample restrictions		
Principals excluded based on either of the following: • The principal was not the only leader of his or her school during a school year • The principal led more than one school during a school year	457	3,812
Calculating principals' contributions to student achievement		
Principals excluded based on not being able to calculate their contribution to student achievement	2,408	1,336
The remaining sample includes:		
Principals who began leading a school during 2008/09, 2009/10, 2011/12, or 2012/13; led the same school the next year; and whose predecessor was in the sample		668
The predecessor principals		668
Obtaining principals' performance ratings		
Principals excluded based on not being able to obtain their performance ratings	1,202	123
The sample for the predictive validity analysis includes:		
Principals who began leading a school in 2008/09 and their predecessor who left in 2007/08, where both in the pair also led a school serving at least one grade in the 5 to 8 range in 2011/12 and 2012/13		33
Principals who began leading a school in 2009/10 and their predecessor who left in 2008/09, where both in the pair also led a school serving at least one grade in the 5 to 8 range in 2012/13 and 2013/14		26
Principals who began leading a school in 2011/12 and their predecessor who left in 2010/11, where both in the pair also led a school serving at least one grade in the 5 to 8 range in 2008/09 and 2009/10		25
Principals who began leading a school in 2012/13 and their predecessor who left in 2011/12, where both in the pair also led a school serving at least one grade in the 5 to 8 range in 2009/10 and 2010/11		50
Total principal-year combinations		134
Total principal-year-subject combinations (the predictive validity analysis sample)		268

Source: Authors' calculations based on data provided by the Pennsylvania Department of Education, 2006/7–2013/14.

Characteristics of principals and students

This section described the characteristics of principals and students in the analysis sample.

Principals. Characteristics of principals are shown in table E2. The first column pertains to all Pennsylvania principals with students in the grade 5 to 8 range in 2010/11, the midpoint of the analysis period. The remaining columns pertain to principals in the predictive validity analysis sample. Specifically, the second column shows the characteristics of incoming principals; the third column shows the characteristics of departing principals. The data in these two columns pertain to the year in which the principal either began or finished leading a school.

The data indicate that incoming principals in the study tended to have fewer years of experience than Pennsylvania principals overall, and were more likely to be racial/ethnic minorities and women. Departing principals, in contrast, tended to have more years of experience, and were less likely to be racial/ethnic minorities and women.

Table E2. Characteristics of Pennsylvania principals, statewide during 2010/11, and for the predictive validity analysis sample during the year in which they entered or departed

		Principals in the predictive validity analysis sample	
Characteristic (percentages unless indicated)	All principals in Pennsylvania 2010/11	Principals in their first year as a school's leader during the validation period (2008/09–2012/13)	Principals in their last year as a school's leader during the validation period (2007/08–2011/12)
Total experience (years)	19.09	17.71	23.67
Highest degree attained			
Bachelor's	15.32	18.18	18.46
Master's	72.63	72.73	69.23
Doctorate	8.02	9.09	9.23
Race and ethnicity			
Black, non-Hispanic	10.81	15.15	4.62
Hispanic	1.42	4.55	1.54
Other race/ethnicity	87.52	80.30	93.85
Gender			
Female	47.32	59.09	43.08
Male	52.56	40.91	56.92
Number of principals	**2,331**	**66**	**65**

Note: Statistics in the table are based on principals leading schools that include any grades from 5 to 8. The number of principals in the last two columns is less than 134 because 3 principals in the predictive validity analysis sample were not leading a school with grades in the 5 to 8 range during 2010/11.

Source: Authors' calculations based on data provided by the Pennsylvania Department of Education, 2007/8–2013/14.

Students. Characteristics of students in grades 5 through 8 during 2010/11, the midpoint of the analysis period, are shown in table E3. The first column pertains to all Pennsylvania students. The second column pertains to students attending schools where the study calculated principals' contributions to student achievement for the predictive validity analysis.

The data in the table indicate that students attending schools in the analysis sample had similar average characteristics to students in Pennsylvania overall in terms of many characteristics, including test scores. Two exceptions were the percentage of students receiving free or reduced-price meals and the percentages of students who were black or Hispanic, which were above average at the schools in the analysis sample.

Table E3. Characteristics of Pennsylvania students during 2010/11, statewide and in schools where the study estimated principals' contributions to student achievement

Characteristic (percentages unless indicated)	Statewide	Students in schools where the study calculated principals' contributions to student achievement for the prediction analysis sample
Math PSSA score (average z-score)	0.00	0.01
Reading PSSA score (average z-score)	–0.01	–0.04
Received free or reduced-price lunch	39.72	45.34
English learner student	2.09	2.72
Received special education	16.19	15.79
Moved schools during school year	7.48	5.00
Overage for grade	0.47	0.26
Male	50.98	49.83
Black, non-Hispanic	13.92	16.38
Hispanic	7.03	8.92
Other race/ethnicity	74.71	70.26
Number of students	505,519	11,887

PSSA is Pennsylvania System of School Assessment.

Note. Statistics in the table are based on students in grades 5–8.

Source: Authors' calculations based on data provided by the Pennsylvania Department of Education, 2007/8–2013/14.

Appendix F. Supplemental findings

This appendix provides additional findings and more detailed versions of findings presented in the report. The first section describes statistical properties of the principal performance ratings, using a broad principal sample. The second section presents detailed findings on the prediction coefficients obtained from various tests of the predictive validity of the principal performance measures. The third section provides findings from tests to assess the potential for bias in the study's validation approach.

Statistical properties of the principal performance ratings

Means and standard deviations for the principal performance ratings according to different measures and samples are shown in table F1. The performance estimates are reported separately for math and reading. The first two columns of data pertain to a broad sample that includes all principals who were in at least their second year as a school's leader. The latter columns pertain to the principal sample used in the predictive validity analysis.

Table F1. Descriptive statistics on the principal performance ratings

	All principals in their second year or later at their school, 2009/10–2013/14		Principals in the predictive validity analysis	
Principal performance measure	Mean	Standard deviation	Mean	Standard deviation
Math				
Average achievement	0.12	0.45	0.08	0.48
Adjusted average achievement	0.02	0.18	0.05	0.19
School value-added	0.03	0.20	0.06	0.22
Adjusted school value-added	0.01	0.19	0.04	0.20
Reading				
Average achievement	0.11	0.43	0.06	0.44
Adjusted average achievement	0.01	0.15	0.00	0.15
School value-added	0.03	0.15	0.03	0.15
Adjusted school value-added	0.01	0.14	0.01	0.14
Number of principal-year combinations	**7,352**		**134**	
Number of principals	**2,424**		**123**	

Source: Authors' calculations based on data provided by the Pennsylvania Department of Education, 2006/7–2013/14.

Means and standard deviations for just the stable parts of the principal performance ratings are shown in table F2, among principals in the predictive validity analysis.

Table F2. Descriptive statistics on the stable parts of the principal performance ratings

	Principals in the predictive validity analysis	
Principal performance measure	Mean	Standard deviation
Math		
Average achievement	0.07	0.41
Adjusted average achievement	[a]	[a]
School value-added	0.02	0.09
Adjusted school value-added	0.01	0.06
Reading		
Average achievement	0.05	0.39
Adjusted average achievement	[a]	[a]
School value-added	0.01	0.06
Adjusted school value-added	0.00	0.04
Number of principal-year combinations	**134**	
Number of principals	**123**	

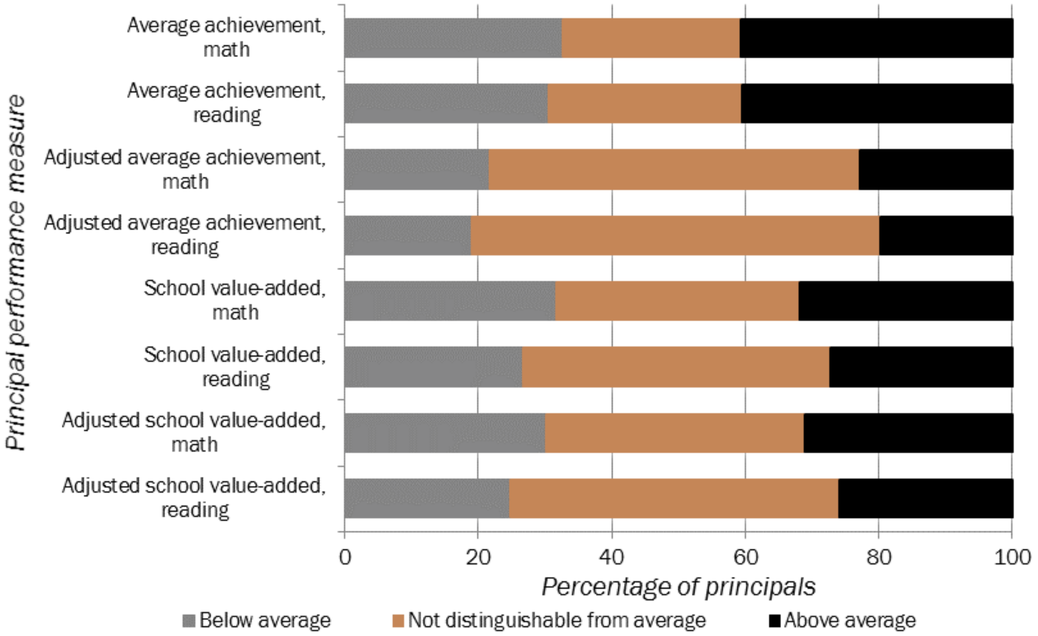
[a] The adjusted average achievement measure had no stable part.

Source: Authors' calculations based on data provided by the Pennsylvania Department of Education, 2006/7—2013/14.

Performance ratings could differ across principals simply by chance. For example, a principal might earn a low performance rating if more students than usual struggled academically in a particular year due to unforeseen illness-related absences. Chance differences do not represent real variation in principal performance. The study confirmed that the measures generated real variation in performance ratings by identifying the percentage of principals whose performance ratings were above or below the average by a statistically significant margin—that is, a margin that should not have been due to chance alone. Across these measures, 20 to 41 percent of principals were statistically distinguishable as above-average performers, and 19 to 33 percent of principals were statistically distinguishable as below-average performers (figure F1).

On three of the four measures, at least half of all principals had single-year performance ratings that were statistically distinguishable from the average. Adjusted average achievement generated the smallest proportion of statistically distinguishable performance ratings, but even on that measure more than one-third of principals differed from the average (44 percent in math and 39 percent in reading). In short, all four measures identified many above- and below-average principals with a high degree of confidence. The ability of each measure to reliably distinguish principals from average is consistent with that of teacher value-added measures examined in the state of Pennsylvania, which can distinguish 30 to 50 percent of teachers from average (Lipscomb, Chiang, & Gill, 2012).

Figure F1. Percentage of principals whose single-year performance ratings were statistically distinguishable or indistinguishable from the average performance



Note: Performance ratings were classified as below average, not distinguishable from average, or above average based on a two-tailed *t*-test at the 5 percent significance level. Figure is based on all principals in Pennsylvania in their second year or later at their school who had performance ratings in any year from 2009/10–2013/14 (N = 7,352 principal-year combinations, N = 2,424 distinct principals).

Source: Authors' calculations based on data provided by the Pennsylvania Department of Education, 2006/7–2013/14.

Correlation coefficients between principal performance estimates from different measures, using only principals in the predictive validity analysis sample, are shown in tables F3 and F4. In both math (table F3) and reading (table F4), each measure was significantly positively related to every other measure. However, the measures did not always agree in their assessments of principals' performance. The correlation between each pair of measures—in which zero would imply no relationship and one would imply that the measures gave identical information about differences across principals—was typically less than 0.7. The exception was that the school value-added and adjusted school value-added measures were highly correlated (0.95 to 0.96).

Table F3. Correlations among performance measures in math for principals in the predictive validity analysis sample

Principal performance measure	Average achievement	Adjusted average achievement	School value-added	Adjusted school value-added
Average achievement	1.00			
Adjusted average achievement	0.52*	1.00		
School value-added	0.65*	0.62*	1.00	
Adjusted school value-added	0.59*	0.65*	0.95*	1.00

* Significant at *p* = 0.05.

Note: Table is based on 134 principal-year combinations and 123 distinct principals in the final prediction analysis.

Source: Authors' calculations based on data provided by the Pennsylvania Department of Education, 2006/7—2013/14.

Table F4. Correlations among performance measures in reading for principals in the predictive validity analysis sample

Principal performance measure	Average achievement	Adjusted average achievement	School value-added	Adjusted school value-added
Average achievement	1.00			
Adjusted average achievement	0.42*	1.00		
School value-added	0.59*	0.55*	1.00	
Adjusted school value-added	0.51*	0.55*	0.96*	1.00

* Significant at *p* = 0.05.

Note: Table is based on 134 principal-year combinations and 123 distinct principals in the final prediction analysis.

Source: Authors' calculations based on data provided by the Pennsylvania Department of Education, 2006/7—2013/14.

Correlation coefficients between principal performance ratings from different measures, using all principals in at least their second year as a school's leader, are shown in table F5 for math and table F6 for reading. The findings for this broad set of principals were similar to those reported in tables F3 and F4, which focused on the principal sample used in the predictive validity analysis. In particular, the findings indicate that all of the principal performance measures were positively related but not the same. The largest correlation was between school value-added and adjusted school value-added.

Table F5. Correlations among performance measures in math for all principals in their second year or later at their school

Principal performance measure	Average achievement	Adjusted average achievement	School value-added	Adjusted school value-added
Average achievement	1.00			
Adjusted average achievement	0.47*	1.00		
School value-added	0.55*	0.54*	1.00	
Adjusted school value-added	0.49*	0.63*	0.95*	1.00

* Significant at *p* = 0.05.

Note: Table includes all principals in Pennsylvania in their second year or later at their school who had performance estimates in any year from 2009/10–2013/14 (N = 7,352 principal-year combinations, N = 2,424 principals).

Source: Authors' calculations based on data provided by the Pennsylvania Department of Education, 2006/7–2013/14.

Table F6. Correlations among performance measures in reading for all principals in their second year or later at their school

Principal performance measure	Average achievement	Adjusted average achievement	School value-added	Adjusted school value-added
Average achievement	1.00			
Adjusted average achievement	0.45*	1.00		
School value-added	0.58*	0.54*	1.00	
Adjusted school value-added	0.50*	0.61*	0.96*	1.00

* Significant at *p* = 0.05.

Note: Table includes all principals in Pennsylvania in their second year or later at their school who had performance estimates in any year from 2009/10–2013/14 (N = 7,352 principal-year combinations, N = 2,424 principals).

Source: Authors' calculations based on data provided by the Pennsylvania Department of Education, 2006/7–2013/14.

Estimates of the prediction coefficient

The next set of tables contains estimates of the prediction coefficient, representing the fraction of the difference in performance ratings between two principals that was reflected in the difference in their contributions to student achievement in a different year.

The main findings from the prediction analysis, reported earlier in figures 6 through 8 of the main report, are shown in table F7. The data reported in the table include findings for both the stable parts of the performance measures and for single-year and three-year rolling averages of the original measures when used to predict contributions to student achievement in the following year. Separate findings for math and reading are presented in table F8 (for the stable part of each measure) and table F9 (for single-year and three-year rolling averages of the original measures). Those tables also include prediction coefficients for an alternative form of adjusted average achievement and adjusted school value-added. The alternative adjustment, which leads to similar results, uses all schools in the state—not just those with leadership transitions, as in the main adjustment method—to estimate the relationship between school performance measures and the same measures two years earlier. None of the findings by subject or for the alternative adjustment were statistically significant.

Table F7. Prediction coefficient estimates from the main findings

Principal performance measure	Prediction coefficient for stable part of the measure	Prediction coefficient for the original measure, when predicting the next year's contribution	Prediction coefficient for a three-year rolling average of the original measure, when predicting the next year's contribution
Average achievement	0.04^ (0.07) {0.56}	0.04^ (0.07) {0.56}	0.04^ (0.07) {0.56}
Adjusted average achievement	[a]	[a]	[a]
School value-added	0.43*^ (0.22) {0.05}	0.29*^ (0.15) {0.05}	0.28*^ (0.14) {0.05}
Adjusted school value-added	0.54 (0.40) {0.18}	0.25^ (0.18) {0.18}	0.25^ (0.18) {0.18}

* Significantly different from 0 at *p* = 0.05.

^ Significantly different from 1 at *p* = 0.05.

[a] The adjusted average achievement measure had no stable part.

Notes: Standard errors are in parentheses. *p*-values for tests of differences from 0 are in brackets. Each coefficient estimate is derived from a separate regression that also includes subject-specific indicators for pairs of outgoing and incoming principals, subject-specific indicators for incoming principals, and school-by-year averages of student background characteristics. The analysis sample consists of 268 principal-year-subject combinations and 123 principals.

Source: Authors' calculations based on data provided by the Pennsylvania Department of Education, 2006/7—2013/14.

Table F8. Prediction coefficient estimates for stable parts of the performance measures: supplementary findings

	Prediction coefficients for stable parts of the measures		
Principal performance measure	Math and reading pooled	Math	Reading
Average achievement			
Unadjusted	0.04^ (0.07) {0.56}	0.03^ (0.08) {0.65}	0.04^ (0.09) {0.63}
Adjusted	[a]	[a]	[a]
School value-added			
Unadjusted	0.43*^ (0.22) {0.05}	0.33^ (0.24) {0.16}	0.69 (0.36) {0.06}
Adjusted	0.54 (0.40) {0.18}	0.46 (0.42) {0.27}	0.74 (0.63) {0.25}
Alternative adjustment	0.54 (0.53) {0.31}	0.53 (0.57) {0.36}	0.66 (0.81) {0.42}

* Significantly different from 0 at *p* = 0.05.

^ Significantly different from 1 at *p* = 0.05.

[a] The adjusted average achievement measure had no stable part.

Table F8. Prediction coefficient estimates for stable parts of the performance measures: supplementary findings (continued)

Notes: Standard errors are in parentheses. *p*-values for tests of differences from zero are in brackets. Each coefficient estimate is derived from a separate regression that also includes subject-specific indicators for pairs of outgoing and incoming principals, subject-specific indicators for incoming principals, and school-by-year averages of student background characteristics. The alternative adjustment uses all schools in the state—not just those with leadership transitions, as in the main adjustment method—to estimate the relationship between school performance measures and the same measures two years earlier. The analysis sample consists of 268 principal-year-subject combinations and 123 principals. Math-only and reading-only columns each use half the sample size (134 principal-year-subject combinations for the 123 principals).

Source: Authors' calculations based on data provided by the Pennsylvania Department of Education, 2006/7—2013/14.

Table F9. Prediction coefficient estimates for the original performance measures: supplementary findings

	Prediction coefficients for original measures, when predicting contributions in the next year			Prediction coefficients for three-year rolling averages of the original measures, when predicting contributions in the next year		
Principal performance measure	Math and reading pooled	Math	Reading	Math and reading pooled	Math	Reading
Average achievement						
Unadjusted	0.04^ (0.07) {0.56}	0.03^ (0.07) {0.65}	0.04^ (0.09) {0.63}	0.04^ (0.07) {0.56}	0.03^ (0.07) {0.65}	0.04^ (0.09) {0.63}
Adjusted	[a]	[a]	[a]	[a]	[a]	[a]
School value-added						
Unadjusted	0.29*^ (0.15) {0.05}	0.22^ (0.16) {0.16}	0.45^ (0.23) {0.06}	0.28*^ (0.14) {0.05}	0.21^ (0.15) {0.16}	0.43^ (0.22) {0.06}
Adjusted	0.25^ (0.18) {0.18}	0.21^ (0.19) {0.27}	0.34^ (0.29) {0.25}	0.25^ (0.18) {0.18}	0.22^ (0.20) {0.27}	0.33^ (0.29) {0.25}
Alternative adjustment	0.25^ (0.25) {0.31}	0.24^ (0.26) {0.36}	0.30^ (0.37) {0.42}	0.25^ (0.25) {0.31}	0.25^ (0.27) {0.36}	0.30^ (0.37) {0.42}

* Significantly different from 0 at $p = 0.05$.

^ Significantly different from 1 at $p = 0.05$.

[a] The adjusted average achievement measure had no stable part.

Notes: Standard errors are in parentheses. *p*-values for tests of differences from zero are in brackets. Each coefficient estimate is derived from a separate regression that also includes subject-specific indicators for pairs of outgoing and incoming principals, subject-specific indicators for incoming principals, and school-by-year averages of student background characteristics. The alternative adjustment uses all schools in the state—not just those with leadership transitions, as in the main adjustment method—to estimate the relationship between school performance measures and the same measures two years earlier. The analysis sample consists of 268 principal-year-subject combinations and 123 principals. Math-only and reading-only columns each use half the sample size (134 principal-year-subject combinations for the 123 principals).

Source: Authors' calculations based on data provided by the Pennsylvania Department of Education, 2006/7—2013/14.

As discussed in appendix D, the benchmark approach to estimating principals' contributions to student achievement relied on principal transitions within schools. In particular, a school's change in average student achievement when one principal replaced another served as the benchmark estimate for the difference between the two principals' contributions. However, prior literature suggests that principal transitions may be correlated with a downward trajectory in performance during the first two years of a new principal's tenure (Miller, 2013). As discussed in appendix D, the estimation model for the prediction coefficients presented in tables F7 through F9 (see equation D1) accounted for the possibility of a systematic dip in achievement around a principal transition that would be constant across all schools. Specifically, the model included subject-specific indicators for being an incoming principal at a school to control for systematic, constant differences in school circumstances between incoming and departing principals.

However, dips in achievement around principal transitions could depend on a school's initial level of student achievement. For example, the dips might be more pronounced for schools with initially higher achievement because there is greater room for a decline. To account for the possibility of dips that are dependent on the school's initial level of student achievement, the study estimated an alternative model specification that interacted the subject-specific indicators for being an incoming principal with the average student achievement in the departing principal's last year at the school. The prediction coefficients from this supplemental analysis are shown in table F10 and are consistent with the findings from the primary analysis reported in table F7.

Table F10. Prediction coefficient estimates from an alternative model specification that accounts for varying dips in school performance during principal transitions

	Prediction coefficients for stable parts of the measures		
Principal performance measure	Math and reading pooled	Math	Reading
Average achievement	0.03^ (0.07) {0.66}	0.02^ (0.08) {0.81}	0.04^ (0.09) {0.66}
Adjusted average achievement	[a]	[a]	[a]
School value-added	0.38^ (0.23) {0.11}	0.27^ (0.26) {0.29}	0.64 (0.34) {0.07}
Adjusted school value-added	0.43 (0.41) {0.30}	0.37 (0.45) {0.42}	0.59 (0.58) {0.31}

^ Significantly different from 1 at $p = 0.05$.

[a] The adjusted average achievement measure had no stable part.

Notes: Standard errors are in parentheses. *p*-values for tests of differences from zero are in brackets. Each coefficient estimate is derived from a separate regression that also includes subject-specific indicators for pairs of outgoing and incoming principals, subject-specific indicators for incoming principals, and school-by-year averages of student background characteristics. Each model also controls for average student achievement at the school in the outgoing principal's last year interacted with the subject-specific indicators for incoming principals. The analysis sample consists of 268 principal-year-subject combinations and 123 principals. Math-only and reading-only columns each use half the sample size (134 principal-year-subject combinations for the 123 principals).

Source: Authors' calculations based on data provided by the Pennsylvania Department of Education, 2006/7—2013/14.

Testing for bias in the validation approach

The study produced evidence about how well four principal performance measures could predict changes in schools' test scores during leadership transitions. The study then interpreted this evidence as indicating how well each performance measure could predict differences in principals' contributions to student achievement. This section discusses the conditions under which this interpretation is valid and the study's approach to assessing threats that could undermine this interpretation.

Requirements for unbiased conclusions about predictive validity. In the predictive validity analysis, changes in schools' test scores during leadership transitions were the study's benchmark estimates of differences between the incoming and departing principals' contributions. However, a school's test scores might rise or fall for any number of factors beyond principals' control, including demographic changes in the school's student population, new district-mandated interventions or other changes in district policies at the school, and unexpected events such as weather-related school cancellations. These factors would lead to errors in the benchmark estimates. In other words, factors beyond principals' control that generated a rise or fall in a school's test scores would lead to an overestimate or underestimate of the successor's contribution relative to that of the predecessor.

The estimated prediction coefficient was unbiased only if errors in the benchmark estimates were not systematically related to principals' performance ratings from the measurement period. That is, the arrival of a successor with a better (or worse) performance rating from the measurement period could not have been systematically related to other changes at the school that increased or decreased test scores.

Formally, the benchmark estimate of the difference in contributions to student achievement between the incoming and departing principal in pair *p* $\left(\Delta\hat{C}_{pk}\right)$ could be expressed as the true difference $\left(\Delta C^{*}_{pk}\right)$ plus estimation error $\left(u_{pk}\right)$:

(F1) $$\Delta\hat{C}_{pk} = \Delta C^{*}_{pk} + u_{pk}\,.$$

The study sought to obtain a true prediction coefficient, β_1^*, that would capture the linear relationship between differences in performance ratings $\left(\Delta\hat{M}_{pk}\right)$ and differences in true contributions to student achievement $\left(\Delta C^{*}_{pk}\right)$:

(F2) $$\beta_1^* = Cov\left(\Delta C^{*}_{pk}, \Delta\hat{M}_{pk}\right) / Var\left(\Delta\hat{M}_{pk}\right).$$

The study actually produced an estimated prediction coefficient, $\hat{\beta}_1$, that captured the linear relationship between differences in performance ratings $\left(\Delta\hat{M}_{pk}\right)$ and the benchmark estimates of differences in contributions to student achievement $\left(\Delta \hat{C}_{pk}\right)$. This estimated prediction coefficient had an expected value of

(F3)

$$E\left(\hat{\beta}_1\right)=\frac{Cov\left(\Delta \hat{C}_{pk},\Delta \hat{M}_{pk}\right)}{Var\left(\Delta \hat{M}_{pk}\right)}=\frac{Cov\left(\Delta C^*_{pk}+u_{pk},\Delta \hat{M}_{pk}\right)}{Var\left(\Delta \hat{M}_{pk}\right)}=\beta^*_1+\frac{Cov\left(u_{pk},\Delta \hat{M}_{pk}\right)}{Var\left(\Delta \hat{M}_{pk}\right)}.$$

The estimated prediction coefficient, $\hat{\beta}_1$, was a biased estimator of β^*_1 if u_{pk} was correlated (that is, had a nonzero covariance) with $\Delta \hat{M}_{pk}$.

Checking for potential sources of bias. To check for bias in the prediction coefficient, the study needed to assess whether the arrival of principals with better or worse performance ratings tended to be accompanied by other changes at their schools beyond their control. Although it was inherently infeasible to examine all types of possible changes, this study checked three key sources of potential bias:

1. **Intensive interventions.** During the period of this study, two types of federal interventions in low-performing schools—restructuring under the No Child Left Behind Act of 2001 and School Improvement Grants (SIGs)—included replacement of the principal as an optional or mandatory component of the intervention.[11] Besides just triggering leadership transitions, those interventions were supposed to bring fundamental changes to the schools' governance structure or instructional approach. If principals with systematically higher (or lower) performance ratings were hired at the onset of such interventions, then the other changes at these schools might have biased the relationship between principals' performance estimates and schools' test scores. For example, if these interventions brought in principals with relatively higher performance ratings, then those principals might have appeared more effective in the validation period than they actually were, due to the beneficial effects of other reforms.
2. **Trends in achievement across cohorts.** If principal transitions coincided with improving student achievement trends (where later cohorts outperformed earlier cohorts upon reaching the same grade level), then the "contributions" of incoming principals would be systematically overestimated. The opposite bias would occur if the later cohorts underperformed the earlier cohorts.

[11] Schools in Pennsylvania that underwent restructuring either had most of their staff replaced, entered into a contract with private operators, or undertook "other major restructuring" (U.S. Department of Education, 2009a, 2010a, 2011, 2012, 2013). The first two types of restructuring actions were likely to have entailed replacing the principal. Schools in Pennsylvania that received SIGs either closed, had a majority of their staff replaced, were converted to a charter school or given to a private operator, or undertook reforms to increase educator effectiveness and learning time. All of the SIG reform models were likely to have entailed replacing the principal (U.S. Department of Education, 2010b).

3. **Changing student characteristics across cohorts.** If principal transitions coincided with changing student characteristics, then principals with the more challenging student populations would have their effects systematically underestimated.

No principals in the predictive validity analysis sample led schools during the validation period that underwent the types of intensive interventions described above. As a result, the presence of intensive school interventions was not a source of bias in the study.

To explore the latter two sources of potential bias, the study used tests known as falsification tests to assess whether differences in performance estimates between incoming and departing principals could predict differences in the prior test scores or characteristics of the students tested under their leadership. The study team estimated versions of equation D1 in which school-by-year average student test scores from two years earlier, gender, race/ethnicity categories, free or reduced-price lunch status, English learner student status, special education status, mobility, and being overage for grade served as dependent variables. These analyses were called falsification tests because the key independent variable in the regression models should not have been able to predict the outcome variables. For instance, the twice-lagged achievement of students present during an incoming principal's second year could not have been affected by the incoming principal. A significant relationship, controlling for other observable measures in the analysis, would raise suspicion that the arrival of principals with better or worse performance ratings might also have been related to unobserved changes at these schools that increased or decreased test scores. This could lead to bias in the prediction coefficient estimates.

The findings from examining whether the stable parts of the four principal performance measures could predict student achievement from two years earlier are shown in table F11. The data suggest that incoming principals with higher or lower performances ratings did not systematically get assigned responsibility for student cohorts with better or worse prior achievement than the preceding principals did.

Table F11. Extent to which the stable part of principal performance measures predicted students' prior achievement

Student characteristic	Stable part of average achievement	Stable part of adjusted average achievement	Stable part of school value-added	Stable part of adjusted school value-added
Achievement from two years earlier	0.07 (0.06) {0.22}	[a]	0.05 (0.20) {0.82}	-0.12 (0.33) {0.72}

[a] The adjusted average achievement measure had no stable part.

Note: Standard errors are indicated in parentheses. *p*-values are indicated in brackets. Each coefficient estimate is derived from a separate regression in which the dependent variable is the school-by-year average of current students' achievement from two years earlier and the independent variable of interest is the performance measure shown in the column. Each regression also includes subject-specific indicators for pairs of outgoing and incoming principals, subject-specific indicators for incoming principals, and school-by-year averages of student background characteristics. The analysis sample consists of 140 principal-year-subject combinations and 66 principals. Principal pairs could not be included if any principal in the pair was missing a value for the twice-lagged achievement variable.

Source: Authors' calculations based on data provided by the Pennsylvania Department of Education, 2006/7—2013/14.

The findings from examining whether the stable parts of the four principal performance measures could predict student background characteristics at the school are shown in table F12. The data reported in this table indicate the potential for bias, based on statistically significant relationships between at least one of the four performance measures and schools' proportions of students who were recipients of free or reduced-price lunches, English learner students, and overage for their grade.

Table F12. Extent to which the stable parts of principal performance measures predicted students' background characteristics

Student characteristic	Stable part of average achievement	Stable part of adjusted average achievement	Stable part of school value-added	Stable part of adjusted school value-added
Male	-0.02 (0.03) {0.40}	[a]	-0.20 (0.11) {0.06}	-0.30 (0.17) {0.08}
Black	-0.01 (0.02) {0.72}	[a]	0.08 (0.05) {0.13}	0.14 (0.09) {0.14}
Hispanic	-0.01 (0.02) {0.68}	[a]	-0.06 (0.05) {0.24}	-0.09 (0.08) {0.24}
Free or reduced-price lunch	-0.04 (0.03) {0.22}	[a]	-0.34* (0.09) {0.00}	-0.64* (0.16) {0.00}
English learner student	0.01 (0.01) {0.45}	[a]	0.10* (0.04) {0.01}	0.19* (0.06) {0.00}
Special education	0.01 (0.02) {0.77}	[a]	0.09 (0.06) {0.11}	0.19 (0.11) {0.08}
Mobile	0.01 (0.00) {0.18}	[a]	0.04 (0.03) {0.13}	0.06 (0.04) {0.12}
Overage for grade	0.00 (0.00) {0.19}	[a]	0.02* (0.01) {0.02}	0.04* (0.02) {0.01}

* Significantly different from 0 at $p = 0.05$.

[a] The adjusted average achievement measure had no stable part.

Note: Standard errors are indicated in parentheses. *p*-values are indicated in brackets. Each coefficient estimate is derived from a separate regression in which the dependent variable is the school-by-year average of the student characteristic shown in the row and the independent variable of interest is the performance measure shown in the column. Each regression also controls for subject-specific indicators for pairs of outgoing and incoming principals, subject-specific indicators for incoming principals, and school-by-year averages of student characteristics other than the dependent variable. The analysis sample consists of 268 principal-year-subject combinations and 123 principals.

Source: Authors' calculations based on data provided by the Pennsylvania Department of Education, 2006/7—2013/14.

To examine the potential for bias due to changing student characteristics across cohorts more closely, the study team restricted the sample in the predictive validity analysis to exclude principals with extreme values of the characteristics where statistically significant relationships had been observed in table F12. This approach effectively reduced the variation across principals in student background characteristics, thereby reducing the risk that performance measures would be related to those characteristics.

Using the restricted sample, the study team did not find any statistically significant relationships for the falsification tests, with three exceptions (tables F13 and F14). The stable part of school value-added was statistically significantly related to schools' proportions of students who were male and overage for their grade. The stable part of adjusted school value-added was statistically significantly related to schools' proportions of students who were overage for their grade. Among those significant relationships, the relationships with the proportion who were overage were always small in magnitude; even a very large difference in value-added ratings of 0.1 student standard deviations (representing more than a standard deviation of ratings across principals) would be associated with a difference of less than one-half of a percentage point in the proportion who were overage. Relationships with proportion who were male were approximately the same magnitude as those in the full sample. Overall, the restricted sample gave rise to few relationships between performance ratings and student background characteristics, but given the remaining small number of relationships, the study could not completely rule out the potential for bias in the prediction coefficients even in the restricted sample.

Table F13. Extent to which the stable parts of the principal performance measures predicted students' prior achievement in the restricted sample

Student characteristic	Stable part of average achievement	Stable part of adjusted average achievement	Stable part of school value-added	Stable part of adjusted school value-added
Achievement from two years earlier	0.05 (0.05) {0.30}	[a]	-0.04 (0.19) {0.84}	-0.24 (0.30) {0.42}

[a] The adjusted average achievement measure had no stable part.

Note: Standard errors are indicated in parentheses. *p*-values are indicated in brackets. This analysis uses only the schools in which the percentage of students who received free or reduced price lunch increased or decreased by no more than 20 percentage points and the percentage of students who were English learner students increased or decreased by no more than 5 percentage points over the first two years of the incoming principal. Each coefficient estimate is derived from a separate regression in which the dependent variable is the school-by-year average of current students' achievement from two years earlier and the independent variable of interest is the performance measure shown in the column. Each regression also includes subject-specific indicators for pairs of outgoing and incoming principals, subject-specific indicators for incoming principals, and school-by-year averages of student background characteristics. The analysis sample consists of 124 principal-year-subject combinations and 59 principals.

Source: Authors' calculations based on data provided by the Pennsylvania Department of Education, 2006/7—2013/14.

Using the restricted sample, the study team then reestimated the prediction coefficient estimates and obtained results qualitatively similar to the main analyses (table F15). Although the data in table F15 do not eliminate the potential for bias in the prediction analyses, the comparability of findings provides reassurance about the robustness of the main results.

Table F14. Extent to which the stable parts of the principal performance measures predicted students' background characteristics in the restricted sample

Student characteristic	Stable part of average achievement	Stable part of adjusted average achievement	Stable part of school value-added	Stable part of adjusted school value-added
Male	-0.03 (0.03) {0.35}	[a]	-0.19* (0.08) {0.03}	-0.23 (0.14) {0.10}
Black	-0.02 (0.02) {0.26}	[a]	0.07 (0.06) {0.30}	0.09 (0.11) {0.39}
Hispanic	-0.01 (0.02) {0.52}	[a]	-0.06 (0.05) {0.28}	-0.09 (0.08) {0.26}
Free or reduced-price lunch	0.02 (0.02) {0.12}	[a]	-0.03 (0.05) {0.57}	-0.09 (0.07) {0.20}
English learner student	0.00 (0.00) {0.35}	[a]	0.03 (0.02) {0.17}	0.05 (0.04) {0.21}
Special education	0.02 (0.02) {0.34}	[a]	0.12 (0.08) {0.13}	0.20 (0.12) {0.11}
Mobile	0.00 (0.01) {0.37}	[a]	0.04 (0.02) {0.12}	0.06 (0.04) {0.15}
Overage for grade	0.00 (0.00) {0.79}	[a]	0.02* (0.01) {0.04}	0.04* (0.02) {0.02}

* Significantly different from 0 at $p = 0.05$.

[a] The adjusted average achievement measure had no stable part.

Note: Standard errors are indicated in parentheses. *p*-values are indicated in brackets. This analysis uses only the schools in which the percentage of students who received free or reduced price lunch increased or decreased by no more than 20 percentage points and the percentage of students who were English learner students increased or decreased by no more than 5 percentage points over the first two years of the incoming principal. Each coefficient estimate is derived from a separate regression in which the dependent variable is the school-by-year average of the student characteristic shown in the row and the independent variable of interest is the performance measure shown in the column. Each regression also controls for subject-specific indicators for pairs of outgoing and incoming principals, subject-specific indicators for incoming principals, and school-by-year averages of student characteristics other than the dependent variable. The analysis sample consists of 224 principal-year-subject combinations and 106 principals.

Source: Authors' calculations based on data provided by the Pennsylvania Department of Education, 2006/7—2013/14.

Table F15. Prediction coefficient estimates from the restricted sample

Principal performance measure	Prediction coefficient for stable part of the measure	Prediction coefficient for the original measure, when predicting the next year's contribution	Prediction coefficient for a three-year rolling average of the original measure, when predicting the next year's contribution
Average achievement	0.10^ (0.08) {0.24}	0.10^ (0.08) {0.24}	0.10^ (0.08) {0.24}
Adjusted average achievement	a	a	a
School value-added	0.56* (0.24) {0.02}	0.37*^ (0.16) {0.02}	0.36*^ (0.15) {0.02}
Adjusted school value-added	0.71 (0.45) {0.12}	0.32^ (0.21) {0.12}	0.32^ (0.21) {0.12}

* Significantly different from 0 at *p* = 0.05.

^ Significantly different from 1 at *p* = 0.05.

[a] The adjusted average achievement measure had no stable part.

Note: Standard errors are in parentheses. *p*-values for tests of differences from 0 are in brackets. This analysis uses only the schools in which the percentage of students who received free or reduced price lunch increased or decreased by no more than 20 percentage points and the percentage of students who were English learner students increased or decreased by no more than 5 percentage points over the first two years of the incoming principal. Each coefficient estimate is derived from a separate regression that includes subject-specific indicators for pairs of outgoing and incoming principals, subject-specific indicators for incoming principals, and school-by-year averages of student background characteristics. The analysis sample consists of 224 principal-year-subject combinations and 106 principals.

Source: Authors' calculations based on data provided by the Pennsylvania Department of Education, 2006/7—2013/14.

Notes

The authors are grateful to staff at the Pennsylvania Department of Education—especially David Volkman, David Ream, and John Nau—for providing the data used in this study. At Mathematica, several individuals provided important contributions to this study. Swaati Bangalore and Claire Postman conducted excellent programming and research assistance, and Clare Wolfendale gave expert feedback on the programming. Steve Glazerman provided thoughtful reviews of early drafts of this report. Jennifer Littel provided editing support, and Kimberly Moody Ruffin handled production.

References

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

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

Cannon, S., Figlio, D., & Sass, T. (2012). *Principal quality and the persistence of school policies* (Working paper). Evanston, IL: Northwestern University.

Chetty, R., Friedman, J., & Rockoff, J. (2014). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review*, *104*(9), 2593–2632.

Chiang, H., Lipscomb, S., & Gill, B. (2016). Is school value added indicative of principal quality? *Education Finance and Policy*, *11*(3), 283–309. http://eric.ed.gov/?id=EJ1106900

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

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

Dhuey, E., & Smith, J. (2013). *How school principals influence student learning* (Working paper). Toronto, ON: University of Toronto. http://eric.ed.gov/?id=ED535648

Dhuey, E, & Smith, J. (2014). How important are school principals in the production of student achievement? *Canadian Journal of Economics*, *47*(2), 634–663.

Goldhaber, D., & Hansen, M. (2008). *Is it just a bad class? Assessing the stability of measured teacher performance* (CRPE Working Paper 2008_5). Seattle, WA: Center on Reinventing Public Education.

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., & Jones, K. (2014, April). *Principal evaluation systems: Current practices and policies*. Paper presented at the meeting of the American Educational Research Association, Philadelphia, PA.

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

Grissom, J. A., Loeb, S., & Master, B. (2013). Effective instructional time use for school leaders: Longitudinal evidence from observations of principals. *Educational Researcher*, *42*(8), 433–444.

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

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). *Have we identified effective teachers? Validating measures of effective teaching using random assignment.* Seattle, WA: Bill & Melinda Gates Foundation, Measures of Effective Teaching Project.

Kane, T. J., & Staiger, D. O. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation* (Working Paper No. 14607). Cambridge, MA: National Bureau of Economic Research.

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

Loeb, S., Kalogrides D., & Béteille, T. (2012). Effective schools: Teacher hiring, assignment, development, and retention. *Education Finance and Policy*, *7*(3), 269–304.

McCaffrey, D., Sass, T., Lockwood, J. R., & Mihaly, K. (2009). The intertemporal stability of teacher effects. *Education Finance and Policy*, *4*(4), 572–606.

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* (Making Connections Report, 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.

Miller, Ashley. (2013). Principal turnover and student achievement. *Economics of Education Review*, *36*, 60-72.

North Carolina Department of Public Instruction. (2013). *Measuring student learning for educator effectiveness: A guide to the use of student growth data in the evaluation of North Carolina teachers.* Retrieved October 2015, from http://www.dpi.state.nc.us/docs/effectiveness-model/student-growth/measuring-growth-guide.pdf.

Purkey, S. C., & Smith, M. S. (1983). Effective schools: A review. *The Elementary School Journal*, *83*(4), 426–452.

SAS Institute, Inc. (2014). *Misconceptions about PVAAS teacher specific reporting for Pennsylvania*. Retrieved October 2015 from http://www.education.pa.gov/Documents/K-12/Assessment%20and%20Accountability/PVAAS/Teacher%20Reports/Misconceptions%20About%20EVAAS%20for%20Teachers.pdf.

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* (Making

Connections Report, 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

U.S. Department of Education. (2009a). *Consolidated state performance report: Part I for state formula grant programs under the Elementary and Secondary Education Act as amended by the No Child Left Behind Act of 2001: For reporting on school year 2008-09, Pennsylvania.* Retrieved October 2015, from http://www2.ed.gov/admins/lead/account/consolidated/sy08-09part1/pa.pdf.

U.S. Department of Education. (2009b). *Race to the Top program executive summary.* Retrieved October 2015, from http://www2.ed.gov/programs/racetothetop/executive-summary.pdf.

U.S. Department of Education. (2010a). *Consolidated state performance report: Part I for state formula grant programs under the Elementary and Secondary Education Act as amended by the No Child Left Behind Act of 2001: For reporting on school year 2009-10, Pennsylvania.* Retrieved October 2015, from http://www2.ed.gov/admins/lead/account/consolidated/sy09-10part1/pa.pdf.

U.S. Department of Education. (2010b). *Guidance on fiscal year 2010 school improvement grants under section 1003(g) of the Elementary and Secondary Education Act of 1965.* Retrieved October 2015, from http://www2.ed.gov/programs/sif/sigguidance11012010.pdf.

U.S. Department of Education. (2011). *Consolidated state performance report: Part I for state formula grant programs under the Elementary and Secondary Education Act as amended by the No Child Left Behind Act of 2001: For reporting on school year 2010-11, Pennsylvania.* Washington, DC.

U.S. Department of Education. (2012). *Consolidated state performance report: Part I for state formula grant programs under the Elementary and Secondary Education Act as amended by the No Child Left Behind Act of 2001: For reporting on school year 2011-12, Pennsylvania.* Retrieved October 2015, from http://www2.ed.gov/admins/lead/account/consolidated/sy11-12part1/pa.pdf.

U.S. Department of Education. (2013). *Consolidated state performance report: Part I for state formula grant programs under the Elementary and Secondary Education Act as amended by the No Child Left Behind Act of 2001: For reporting on school year 2012-13, Pennsylvania.* Retrieved October 2015, from http://www2.ed.gov/admins/lead/account/consolidated/sy12-13part1/pa.pdf.