July 2014

Professional practice, student surveys, and value-added: Multiple measures of teacher effectiveness in the Pittsburgh Public Schools

Duncan Chaplin
Brian Gill
Allison Thompkins
Hannah Miller
Mathematica Policy Research

Key findings

- All three measures of teacher effectiveness being developed by the Pittsburgh Public Schools—professional practice, student surveys, and value-added measures—have the potential to differentiate teacher performance.
- The measures are positively, if moderately, correlated, suggesting that they are valid and complementary measures of teacher effectiveness.
- Variation across schools on the professional practice rating (which is assigned to teachers by the school principal) suggests that principals' standards might not be fully consistent across schools and that the measure might be improved by using more than one rater for each teacher.

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

REL MID-ATLANTIC
Regional Educational Laboratory
At ICF International

REL 2014–024

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.

July 2014

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

Chaplin, D., Gill, B,. Thompkins, A., & Miller, H. (2014). *Professional practice, student surveys, and value-added: Multiple measures of teacher effectiveness in the Pittsburgh Public Schools* (REL 2014–024). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Mid-Atlantic. Retrieved from http://ies.ed.gov/ncee/edlabs.

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

Summary

Responding to federal and state prompting, school districts across the country are implementing new teacher evaluation systems that aim to increase the rigor of evaluation ratings, better differentiate effective teaching, and support personnel and staff development initiatives that promote teacher effectiveness and ultimately improve student achievement. States and districts are implementing richer measures of professional practice alongside "value-added" measures of student achievement growth and in some cases are incorporating additional measures, such as student surveys.

Pittsburgh is a leader in the nationwide movement to evaluate, enhance, and reward effective teaching. The analyses presented in this report were conducted to assist Pittsburgh Public Schools in refining its multiple measures of teacher effectiveness, to create a rich, valid, and comprehensive combined measure. The analyses help to assess how well Pittsburgh's measures differentiate among teachers and to establish how strongly they are correlated with each other. In addition, because each of Pittsburgh's three primary measures of teacher effectiveness is based on an approach that is being used or considered elsewhere, the findings have important implications for districts and states across the country that are developing and refining their measures of teacher effectiveness.

The Pittsburgh Public Schools teacher evaluation system includes three types of multicomponent measures. The first measure—the Research-based Inclusive System of Evaluation (RISE), based on Charlotte Danielson's widely used Framework for Teaching (Danielson, 2013)—is an observation-based professional practice measure that relies on principals' assessments. The second measure is based on a student survey called the 7Cs, which incorporates students' perceptions of teachers' practices and was developed by Ronald Ferguson of Harvard University as part of the Tripod Project and administered by Cambridge Education (Bill & Melinda Gates Foundation, 2012). The third measure is a value-added measure that uses changes in student test scores to estimate each teacher's contribution to student achievement over up to three years of teaching (Johnson, Lipscomb, Gill, Booker, & Bruch, 2012). This study used 2011/12 data to describe how the ratings on the three measures are distributed across teachers and how the ratings are correlated.

Key findings

All three measures have the potential to differentiate among teachers. While all three composite measures show a wide range of teacher effectiveness, only the district's value-added measures have been shown to reliably differentiate among teachers (Johnson et al., 2012); the reliability of the RISE and 7Cs composites cannot be determined without multiple ratings per teacher (ideally by multiple raters). However, the components of each of these measures are highly correlated, indicating that the composites have acceptable levels of internal consistency.

Correlations among measures suggest they are valid and complementary. The composites of all three measures are also positively correlated. Teachers with high RISE ratings tend to have high 7Cs ratings and high value-added measure estimates as well. The correlations are moderate but statistically significant—consistent with other research on similar measures of teacher effectiveness (table S1). These results suggest that the measures

Table S1. Correlations of three measures of teacher effectiveness in Pittsburgh Public Schools, 2011/12

Measure	RISE composite	7Cs composite
7Cs composite	.30 (887 teachers)	na
Value-added estimate	.20 (358 teachers)	.15 (619 teachers)

na is not applicable. RISE is Research-based Inclusive System of Evaluation. 7Cs are student survey measures.

Note: All correlations are statistically significant at the 0.05 level.

Source: RISE and 7Cs data are from Pittsburgh Public Schools for the 2011/12 school year; value-added measure estimates are from Johnson et al. (2012) and are for 2008–11.

capture teaching skills that overlap but are not identical—as the district intended in creating multiple measures.

The fact that RISE ratings are significantly correlated with 7Cs and value-added measures despite relying on a single rater (the principal) and despite excluding many of the lowest performing teachers from the measurements (to meet older evaluation requirements that were still in place in 2011/12 but have since been fully replaced with RISE) is a credit to the RISE rating system, which involves extensive training and systematic, year-long data gathering in addition to classroom observations.

Using multiple raters may improve the reliability and validity of RISE. For all three measures, most of the variation occurs within rather than between schools. This is consonant with the literature on value-added measures, which consistently finds larger variation in estimated teacher effectiveness within than between schools. Nonetheless, systematic differences in RISE ratings remain by school even after accounting for differences in value-added and 7Cs measures, suggesting that some principals are tougher or more lenient than others in applying RISE. Using an additional rater for each teacher could help principals better calibrate their RISE ratings, thus enhancing the consistency, fairness, and validity of the ratings (particularly if the additional raters work in more than one school).

In addition, previous research has demonstrated that using more than one rater for each teacher improves the reliability of ratings of professional practice (Kane & Staiger, 2012). If teachers view the additional raters as peers, this might enhance the face validity of the measure and increase teacher buy-in.

Pittsburgh Public Schools, reflecting its interest in continuously improving RISE, assigned instructional teacher leaders to take on some of the classroom observations for at least some teachers beginning in the 2012/13 school year. But maximizing the reliability and validity of RISE ratings requires a systematic effort to produce independent ratings by having two raters for every teacher and having some raters work in multiple schools.

Findings support using all three measures for evaluation

Both RISE and 7Cs, like value-added measures, have the potential to increase differentiation among teachers beyond that achieved through traditional evaluation measures. Although none of the measures represents a gold-standard benchmark, the correlations across them suggest that they are capturing various aspects of effective teaching in complementary ways. Using multiple raters for RISE has the potential to further enhance the evaluation system, which is already producing richer and more fine-grained information on teacher effectiveness than has previously been available in Pittsburgh schools.

Contents

Box

Figures

Tables

Why this study?

As part of a larger effort to develop and reward effective teachers, school systems across the country have been reforming their teacher evaluation systems, with active federal and state encouragement. Pittsburgh Public Schools is a leader in this endeavor and has received support through grants from the federal Teacher Incentive Fund program and the Bill & Melinda Gates Foundation. Working collaboratively with the Pittsburgh Federation of Teachers, the Pittsburgh school district has been developing more-intensive professional development opportunities, a modified career ladder, financial awards for highly effective teaching, and several new multicomponent measures of teacher effectiveness. In addition, the district has begun implementing multiple new measures of teacher effectiveness, with the aim of producing a rich, valid, and comprehensive combined measure. As a member of the Teacher Evaluation Research Alliance of the Regional Educational Laboratory (REL) Mid-Atlantic, Pittsburgh Public Schools asked REL Mid-Atlantic to assist in further developing and analyzing its teacher evaluation measures.

Pittsburgh Public Schools has begun implementing multiple new measures of teacher effectiveness, with the aim of producing a rich, valid, and comprehensive combined measure

All three of Pittsburgh's new measures of teacher effectiveness—related to professional practice, student achievement growth, and student surveys—are based on measures that are in use or under consideration in many districts and states. Thus, the findings of this study should be useful not only to educators in Pittsburgh but to others across the country who are developing new systems of teacher evaluation.

Pittsburgh's new measures include the following (see appendix A for more detail):

- *An observational measure of teacher practice, known as the Research-based Inclusive System of Evaluation (RISE),* which is based on evaluations by the school principal and draws on Charlotte Danielson's Framework for Teaching (Danielson, 2013). Designed to be richer, more rigorous, more differentiated, and more useful for instruction improvement than traditional measures of professional practice, it aims to better differentiate effective teaching and better identify the professional development needs of individual teachers. Feedback is provided to teachers in meetings with principals (the raters in the RISE system) to discuss the ratings they receive, teaching practices that need to be strengthened, and guidance for improvement. Pittsburgh rolled out RISE in phases between 2009 and 2011.

  RISE includes 24 components distributed among four domains (planning and preparation, classroom environment, teaching and learning, and professional responsibilities; see table A1 in appendix A). For the 2011/12 school year the district identified 12 of the 24 as "power components" thought to have the greatest impact on student learning and growth. Principals categorize teachers' performance on each RISE component as unsatisfactory (0), basic (1), proficient (2), or distinguished (3). These ratings are averaged to create composite ratings. Most teachers also receive an overall RISE score that is based on the principal's judgment rather than an average of the component ratings.
- *A measure of students' perceptions of teacher practices,* using the 7Cs surveys developed by Ron Ferguson of Harvard University and administered by Cambridge Education (Bill & Melinda Gates Foundation, 2012). Students assess teachers on survey questions in seven areas (caring for students, controlling the classroom, clarifying the message, challenging learners, captivating the classroom, conferring with students, and consolidating lessons learned). The district began to use 7Cs during the 2011/12 school year.

- *A value-added measure* (VAM), developed in collaboration with Mathematica Policy Research and based on the year-to-year test score growth of individual students. This report employs the teacher VAM estimates calculated by Johnson et al. (2012) using Pittsburgh Public Schools' longitudinal student-level data. VAM estimates were calculated by grade (4–12), subject (math, reading/English language arts, social studies, and science), and test type (Pennsylvania System of School Assessment and curriculum-based assessments) for 2008–12.

The expectation is that together the three new measures will yield a valid and reliable overall gauge of teacher effectiveness. In addition, the district aims to use the measures of teacher effectiveness to provide data that inform professional development and improve teaching practice for individual teachers and the district overall.

What the study examined

The study has two major aims:

- To determine whether the variation across the measures is large enough to be useful for evaluating teachers. Traditional teacher evaluation measures have been criticized for failing to distinguish among the great majority of teachers. Pittsburgh believes that the improvement of teaching practice requires the ability to identify high-performing teachers as well as low-performing teachers.
- To determine the correlations of RISE (professional practice) and 7Cs (student survey) measures with student achievement growth as captured by a teacher's estimated value added. Creation of a combined measure of teacher effectiveness is premised on the assumption that each of the three measures captures a different aspect of a teacher's underlying effectiveness. If so, ratings on the three measures should be positively correlated (teachers who do well on one measure should to do well on the others).

The study aims to determine whether the variation across the measures is large enough to be useful for evaluating teachers and to determine the correlations of RISE and 7Cs measures with student achievement growth as captured by a teacher's estimated value added

The study includes both descriptive analyses (addressing the first aim) and correlational analyses (addressing the second aim). The correlational analyses assess the extent to which teachers' ratings on RISE and 7Cs measures (including composite ratings and ratings of individual components) are related to each other and to student achievement growth as captured by the VAM. (See box 1 for a discussion of the study data and methods.)

Descriptive questions

1. How much variation is evident across teachers on RISE measures of professional practice and 7Cs student survey measures? How are the RISE and 7Cs ratings distributed (how many teachers earn the most common ratings, and are ratings skewed toward the high or low end of the scale)?
2. Can a simple average of RISE components usefully summarize variation across components and domains? This is important to assess in order to know whether the district might need a more complex method of combining RISE components into a composite measure
3. Is the variation in RISE and 7Cs ratings related to observable characteristics of students and teachers? If student characteristics are related to RISE or 7Cs measures, the district might want to consider adjusting the measures to account for those characteristics.

4. Does the variation in RISE and 7Cs ratings occur largely between or within schools? If between schools, does the evidence suggest that some raters are systematically more lenient or more rigorous than others in applying the RISE rubric?

Correlational questions

5. To what extent are teachers' RISE observational ratings (components and composite) correlated with estimates of teachers' value-added contributions to their students' achievement growth?
6. To what extent are teachers' RISE observational ratings (components and composite) correlated with student perceptions as measured by the seven 7Cs ratings (components and composite)?
7. To what extent are teachers' 7Cs ratings (components and composite) correlated with teachers' value-added contributions to their students' achievement growth?

Because none of the individual measures is comprehensive, Pittsburgh Public Schools is seeking to combine the results from RISE observations, student 7Cs surveys, and VAM estimates in a global composite measure that yields a richer picture of teacher effectiveness. But complementary does not mean uncorrelated. The district expects to see significant positive correlations because the RISE and 7Cs measures cover elements of classroom practice that are presumed to be associated with gains in student achievement on state assessments (as measured by the VAMs).

Establishing correlations among measures does not demonstrate causality, but it can provide valuable evidence for educators and researchers on the elements of teaching that are associated with improvements in student achievement. The analysis can also identify ways to simplify and improve the usefulness of the RISE observational metric and can inform efforts to rate teachers lacking VAM estimates by using only RISE and 7Cs ratings. Finally, the analysis can guide the district's efforts to create a composite measure of teacher effectiveness that includes several component measures.

Overview of study findings

The study's findings suggest that, taken together, RISE, 7Cs, and VAM ratings are useful and complementary measures of teacher effectiveness.

RISE and 7Cs composite ratings are internally consistent and have the potential to distinguish differences among teachers' performance

For each RISE observational component, a majority of teachers earned a "proficient" rating in 2011/12. Although each component can have only one of four values, averaging the components into a composite rating yields many more values. The composites that incorporate 12 or 24 RISE components have high levels of internal consistency, suggesting that they capture a coherent underlying concept of teacher effectiveness. However, having more than one year of data would help determine the degree to which RISE measures true differences in performance.

The 7Cs student survey component ratings are reported in normal curve equivalent units, which are designed to create a classic bell-shaped distribution of ratings. As a consequence, all 7Cs component ratings have similar standard deviations, showing considerable variance

Box 1. Data and methods

The analyses in this report use unique teacher IDs to link Pittsburgh Public Schools' three data sources on teacher effectiveness (Research-based Inclusive System of Evaluation [RISE], 7Cs surveys, and value-added measures [VAMs]), focusing on the 2011/12 school year. The analyses also incorporate VAM data covering up to three previous years of performance.

Data. Pittsburgh Public Schools collects data on the 12 RISE power components for a large fraction of its teachers and data on 12 additional components for a smaller fraction of teachers (table B2 in appendix B). The number of RISE components available for each teacher in a given year varied from 0 to 24, plus an overall RISE rating. Approximately a third of teachers each year are not given RISE ratings while they focus on a single component. Teachers who were on improvement plans due to unsatisfactory ratings in prior years were not given RISE ratings in 2011/12. Nearly all of the other 1,071 teachers had ratings for the 12 RISE power components.

The analyses of 7Cs use data for classrooms with at least five completed 7Cs surveys to avoid including teachers with very noisy measures. Eighty-four percent of teachers with at least 12 RISE components also have 7Cs ratings from at least five students.

The VAM data cover a larger percentage of teachers in Pittsburgh Public Schools than in many other school districts, because Pittsburgh schools use not only the Pennsylvania System of School Assessment in reading and math in grades 3–8 but also locally developed curriculum-based assessments tied to specific courses (reading, math, science, and social studies) in secondary grades. The methods used to estimate the VAMs are described in depth in Johnson et al. (2012). The data include VAM estimates for more than 700 teachers (about a third of teachers in the district) and are based on 39 assessment instruments (table A2 in appendix A).

Reliability. A measure's reliability describes how well differences among teachers' ratings at one point in time reflect true differences in teachers' skills that would also be observed at a different time, by a different rater, or with a different measurement instrument. There are several types of reliability. First, test-retest reliability measures the extent to which a teacher would earn the same rating from the same rater if the rater observed the teacher at other times. Second, inter-rater reliability measures the extent to which two raters observing the same lesson would assign the teacher the same rating. These two forms of reliability require multiple observations of the same teacher (not the case here). Only one observation for each teacher is needed for a third form of reliability, internal consistency, which measures the similarity of ratings across different components of teacher effectiveness. Cronbach's alpha is a measure of internal consistency that ranges from zero to one. A value above 0.80 is usually considered sufficient for screening purposes (Wasserman & Bracken, 2003); an alpha above 0.70 is considered acceptable (de Vaus, 2002).

Analyses. Most of the analyses are limited to teachers who met three criteria: teaching at one of the 58 regular schools in the district; teaching math, science, social studies, or English language arts; and rated on at least 12 RISE components. During the 2011/12 school year a third of teachers participated in a supported growth project that involved an intensive focus on a single RISE component. These teachers were rated on only one component and so were excluded from most of the analyses. In addition, about 5–7 percent of teachers were working on employee improvement plans as a result of prior unsatisfactory evaluations. Most of these

(continued)

Box 1. Data and methods *(continued)*

teachers did not participate in the RISE process because the school system did not want to apply high-stakes negative consequences to RISE ratings as the system was being launched. After the 2011/12 school year the employee improvement plan process was folded into RISE. The analyses included a total of 2,082 teachers. However, not all teachers had complete data: only 329 teachers had 7Cs data, data on at least 12 RISE components, and VAM data (tables B1 and B2 in appendix B).

The results are based on simple descriptive statistics, analysis of variance, ordinary least squares regressions, and correlations.[1] Exceptions are the VAM estimates, which are from Johnson et al. (2012), and the results in appendix C, which describes summaries calculated using principal components analysis. All analyses are unweighted unless stated otherwise. Measures reported in the main report are based on averages. Appendix D describes various alternative ways of creating composites of the RISE components. Appendix E examines within- and between-school variation in RISE ratings to help determine whether principals are giving consistent ratings across schools.

[1] Standard errors were not adjusted for clustering. The correlations are Pearson correlations, which use the original scales, as opposed to Spearman correlations, which are based on ranks.

across the scale range. Averaging each teacher's score across 7Cs components concentrates the distribution, so that most teachers have average 7Cs ratings that are close to the districtwide average. Nonetheless, variation remains substantial, and the 7Cs average has a high level of internal consistency.

Because none of the individual measures is comprehensive, Pittsburgh Public Schools is seeking to combine the results from three measures in a global composite measure that yields a richer picture of teacher effectiveness

A simple average of RISE power components provides a useful composite measure of variation across RISE domains and components

A principal components analysis identifies just four variables based on weighted averages of the 12 RISE power components that explain more than half the variation in the power components (appendix C). The first of these variables, which includes all 12 RISE power components, uses weights that differ only slightly. As a consequence, a RISE composite consisting of the 12 equally weighted power components produces nearly identical results. There is no evidence that ratings on the additional 12 components (available for a small fraction of teachers) provide substantially more information about teacher performance than that provided by the 12 power components.

RISE and 7Cs ratings are associated with some student characteristics

Both RISE and 7Cs ratings are correlated in similar ways with student characteristics. Teachers with more low-income and racial/ethnic minority students tend to receive lower ratings on RISE and 7Cs, while teachers with more gifted students tend to receive higher ratings. These differences have two possible interpretations: that lower income and nongifted students are assigned lower performing teachers or that the RISE and 7Cs measures are biased against teachers with more low-income students and nongifted students. Understanding which interpretation is correct would require data on how students with different characteristics experience the same teachers. This could be examined in future studies using student-level 7Cs and RISE data for multiple years.

RISE ratings are correlated with 7Cs ratings and value-added measure estimates, but some principals may be easier raters than others

Within schools, principals' RISE ratings of teachers were significantly (if moderately) correlated with students' 7Cs survey responses on teachers and with teachers' VAM estimates. Principals' ratings were also significantly (moderately) correlated with VAM estimates across schools. These findings help validate principals' ability to rate teachers using the RISE system. Still, substantial variation in RISE ratings between schools remained even after controlling for 7Cs ratings and VAM estimates. It could be that teachers in different schools differ on dimensions that are captured in RISE ratings but not captured in the 7Cs or VAM measures. But it seems at least as likely that systematic differences in RISE ratings across schools, after accounting for 7Cs ratings and VAM estimates, are the result of systematic differences in principals' calibration of RISE standards.

Within schools, principals' RISE ratings of teachers were significantly (if moderately) correlated with students' 7Cs survey responses on teachers and with teachers' VAM estimates

All three measures are positively correlated with each other

The data show significant (if moderate) positive correlations between RISE ratings—especially the RISE composite—and VAM estimates. RISE correlations are somewhat stronger for the VAM estimates for reading than for the other VAM estimates.

The data show very strong correlations between RISE and 7Cs ratings. About 90 percent of the correlations between the components of the RISE and 7Cs measures are statistically significant at the .05 level, and the RISE composite correlates positively and significantly with every one of the 7Cs components.

The data show positive and statistically significant correlations between 7Cs ratings (components and composite) and the VAM estimates, especially in math and at the high school level. Among the 7Cs components, "controls" has the most consistent positive relationship with the VAM estimates.

Distributions of ratings on each teacher evaluation measure

This section presents a more detailed discussion of the descriptive characteristics of the RISE ratings, 7Cs ratings, and VAM estimates, answering the first four research questions. It looks at the distribution of ratings and their reliability. RISE receives the most attention, because it has not been as extensively examined in prior research as the other two measures.

Average RISE ratings are distributed broadly among teachers, and its components are internally consistent

In principle, the four-point scale for the RISE components (0 unsatisfactory, 1 basic, 2 proficient, and 3 distinguished) allows more differentiation of teacher performance than do traditional measures that use a simple satisfactory/unsatisfactory rating. But whether such differentiation occurs depends on how raters use the metric. Kane and Staiger (2012) found that the Danielson Framework for Teaching can be used to produce ratings that vary substantially across teachers. But a recent pilot of a Framework for Teaching–based observation metric conducted for the Pennsylvania Department of Education found that some principals gave nearly all of their teachers identical ratings (Lipscomb, Chiang, & Gill, 2012).

For each component, “proficient” (2) was the most common 2011/12 RISE rating. The percentage of teachers receiving a proficient rating ranged from 57 percent for one component to 88 percent for another (table B3 in appendix B). “Unsatisfactory” (0) was the least common rating, given to no more than 1 percent of teachers. On the overall RISE rating given by principals (as distinguished from the composite rating derived from the components), 86 percent of teachers were rated proficient, with 9 percent rated basic (1) and 5 percent rated distinguished (3).

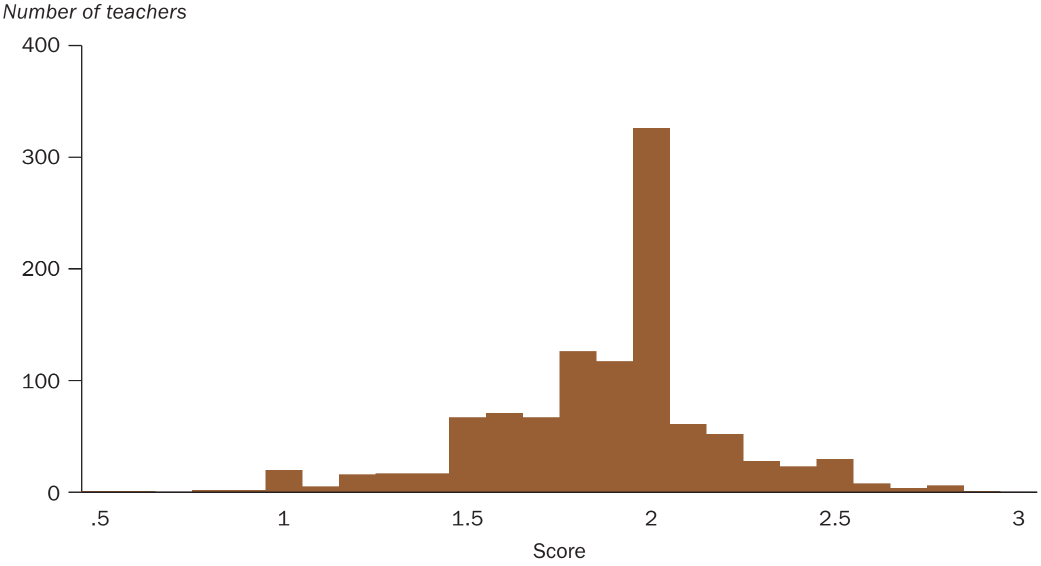
The low percentage of teachers receiving an unsatisfactory rating may be due in part to the omission in the data of teachers who were very low performers during the prior school year. These teachers (approximately 80) were put on improvement plans and not given RISE ratings during the 2011/12 school year.

For each component, “proficient” was the most common 2011/12 RISE rating, with the percentage of teachers receiving a proficient rating ranging from 57 percent for one component to 88 percent for another

Combining the component ratings to create composites can increase the number of possible scores teachers can receive (figure 1). The district intends to use a multicomponent average in its comprehensive teacher evaluation measure. One method of combining the RISE components is to use a simple average of the 12 power components and the overall RISE rating. The distribution of this composite rating shows that it can more finely distinguish the performance of teachers than can the component measures, each of which has only four possible values.

Even though the variation in the RISE ratings, especially the composite ratings, suggests the potential to differentiate the effectiveness of individual teachers from average effectiveness, that does not necessarily mean that teacher effectiveness can be differentiated in a statistically reliable manner. A measure’s reliability describes how well differences

Figure 1. Combining the component RISE ratings to create composites can increase the number of possible scores (distribution of composite based on average of RISE power components)



RISE is Research-based Inclusive System of Evaluation.

Note: RISE ratings are on a scale of 0 to 3, with 3 the highest. The composite score is the average across the 12 power components. The total sample size is 1,068.

Source: Pittsburgh Public Schools data for the 2011/12 school year.

between teachers' observed ratings at one point in time reflect true differences in teacher skills that would also be observed at a different time, by a different rater, or with a different measurement instrument.

Because the RISE data include only one rater and one year for each teacher, internal consistency is the only one of the three types of reliability (see box 1) that can (as yet) be measured. The RISE composite rating (here an average of the 12 power component ratings) has a Cronbach's alpha of 0.87 (a value above 0.80 is considered sufficient for screening purposes; Wasserman & Bracken, 2003).

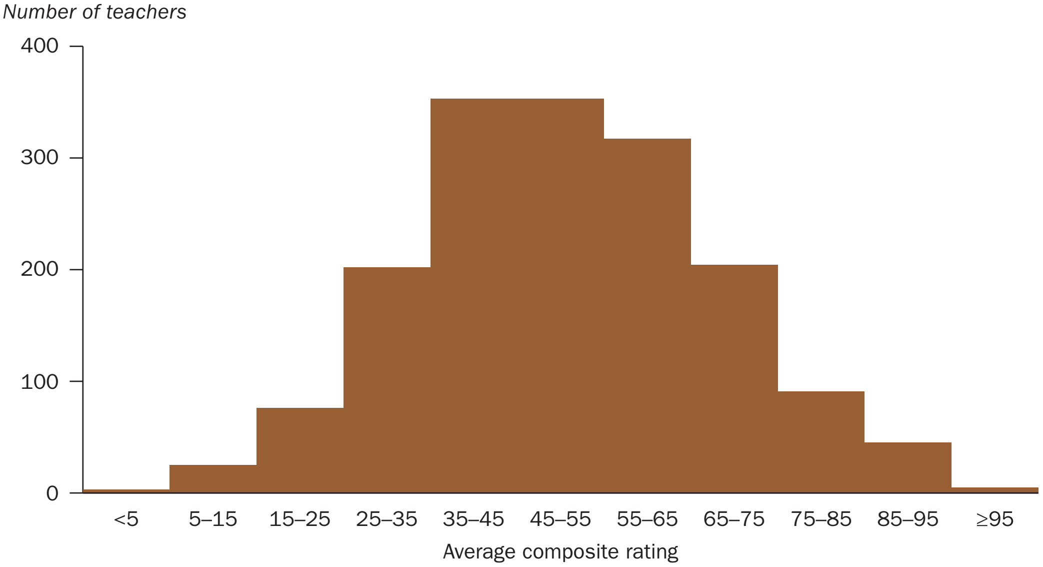
By treating each RISE component as a separate observation of a teacher's general effectiveness, it is possible to estimate what fraction of teachers stand out as particularly high or low performers, given their raters.[1] These estimates of precision account for variation across components but not for variation related to the rater or the day of observation. Under this approach about 39 percent of teachers differ statistically from the mean based on the composite of the 12 RISE power component scores. This probably somewhat overestimates the reliability and precision of the RISE estimates because it cannot account for inter-rater or test-retest reliability.

Estimates based on average 7Cs ratings could distinguish the performance of 65 percent of teachers from the average

7Cs average ratings are distributed broadly among teachers and are internally consistent

Other school districts have used 7Cs survey measures to distinguish teachers who are highly rated and those who are poorly rated by their students (Bill & Melinda Gates Foundation, 2010; Kane, 2012). In Pittsburgh, the distribution of composite 7Cs ratings (unweighted averages of the component scores), grouped into bins, demonstrates the expected spread between 1 and 99 (figure 2), and the standard deviations of the component measures are all between 20 and 21 by design (table B5 in appendix B), since they are normal-curve-equivalent ratings (see appendix A for details).

Figure 2. The distribution of teacher-level composite 7Cs ratings, grouped into bins, demonstrates the expected spread between 1 and 99



Note: 7Cs data are based on student surveys. Values at the cutpoints fall in the higher bin. The total sample size is 1,674.

Source: Pittsburgh Public Schools data for the 2011/12 school year.

As with RISE, the availability of only a single set of ratings for each teacher (in the absence of student-specific responses) meant that it was possible to measure internal consistency but not other elements of reliability.[2] The average 7Cs rating is internally consistent, with a Cronbach's alpha of 0.94. Estimates based on average 7Cs ratings could distinguish the performance of 65 percent of teachers from the average. Again, this probably overestimates the measure's true reliability because it does not account for variance across students.

Value-added measure estimates are distributed broadly and are sufficiently reliable to distinguish teacher performance at the high and low ends of the scale

On average across all VAM estimates, the performance of a third of teachers could be reliably distinguished from the district average (table 1). These results varied by grade, test, and subject type (Pennsylvania System of School Assessment and curriculum-based assessments; Johnson et al., 2012). VAM estimates were somewhat less reliable for the curriculum-based assessments, as might be expected of home-grown assessments, but most of these VAMs could nonetheless statistically distinguish teachers at the high and low ends of the scale. The standard deviation of teacher effects across all assessments averaged 0.26, similar to the results reported in other studies of VAMs (Lipscomb et al., 2012). The VAM data are derived from achievement growth data for many students for each teacher, permitting reliability and precision to be measured based on variation across students.

On average across all VAM estimates, the performance of a third of teachers could be reliably distinguished from the district average

Scores for all three measures show more variation in teacher effectiveness within schools than between schools, consistent with other research

At least 85 percent of the variation in composite ratings of RISE, 7Cs, and VAMs in Pittsburgh Public Schools is within schools; 15 percent or less is between schools (table 2). These findings are consistent with the results for just about every school district where the issue has been examined (Lipscomb et al., 2012). This result implies that there are substantial numbers of effective teachers in low-achieving schools and ineffective teachers in high-achieving schools.

The RISE composite based on 12 power components correlates highly with other possible composites

Principal components analysis was used to develop a composite measure of the RISE power components (see appendix C). Because the resultant composite measure is very similar to an unweighted average of the RISE power components, little is gained by using a composite based on weights identified in the principal components analysis (appendix C). Further, the statistical properties of a composite derived from principal components analysis are only slightly better than those of a composite based on an unweighted average of power components, and that slight improvement comes at the cost of a loss in transparency of the measure.

For these reasons analyses focused on the composite that is an unweighted (or equally weighted) average of the 12 power components. Implicitly, this gives a 17 percent weight to the planning and preparation domain (with two power components), 17 percent to the classroom environment domain (two power components), 42 percent to the teaching and learning domain (five power components), and 25 percent to the professional responsibilities domain (three power components). The unweighted power component average

Table 1. Results for teacher value-added measure estimates, by grade and outcome, 2008–11

Grade	Outcome	Number of teachers	Standard deviation of teacher effects	Mean standard error	Percent statistically significant[a]
4	PSSA Math	80	0.21	0.10	34
	PSSA Reading	86	0.17	0.10	24
	PSSA Science	50	0.24	0.09	46
5	PSSA Math	69	0.21	0.10	32
	PSSA Reading	84	0.15	0.10	27
	PSSA Writing	73	0.28	0.12	44
6	PSSA Math	78	0.18	0.09	32
	PSSA Reading	98	0.17	0.10	31
	CBA Math	54	0.40	0.15	39
	CBA English	63	0.20	0.12	21
	CBA Earth Science	48	0.41	0.11	69
7	PSSA Math	72	0.14	0.08	32
	PSSA Reading	92	0.08	0.08	10
	CBA Math	53	0.28	0.14	30
	CBA English	66	0.14	0.11	14
	CBA Life Science	37	0.35	0.09	59
8	PSSA Math	62	0.18	0.08	39
	PSSA Reading	75	0.17	0.09	29
	PSSA Science	37	0.14	0.07	32
	PSSA Writing	74	0.23	0.12	27
	CBA Math	32	0.53	0.19	41
	CBA English	54	0.26	0.11	35
	CBA Physics	31	0.46	0.11	48
	CBA US History	31	0.41	0.17	52
9	CBA Algebra I/AB-BC	50	0.42	0.16	38
	CBA ELA I	42	0.26	0.14	26
	CBA Biology	23	0.34	0.13	52
	CBA Civics	30	0.34	0.10	60
10	CBA Geometry/AB-BC	31	0.28	0.13	32
	CBA ELA II	27	0.15	0.13	11
	CBA Chemistry	21	0.27	0.15	24
	CBA World History	23	0.37	0.18	43
11	CBA Algebra II	32	0.51	0.21	50
	CBA ELA III	30	0.28	0.17	20
	CBA Physics	14	0.39	0.27	21
	CBA US History	19	0.33	0.15	32
	PSSA Reading	38	0.05	0.08	0
	PSSA Writing	37	0.15	0.14	8
12	CBA ELA IV/AA Lit	25	0.13	0.15	4
All	Average across outcomes[b]	700	0.26	0.13	33

PSSA is Pennsylvania System of School Assessment. CBA is curriculum-based assessment. ELA is English language arts.

Note: The results are based on student z-score units, calculated by grade, subject, and assessment. They are equal to the student scale score, minus the mean and divided by the standard deviation.

a. The percentage of estimated teacher effects that were statistically different from 0 at the 0.05 level.

b. A simple average across the rows of the table.

Source: Johnson et al. (2012).

Table 2. Percentage of variation within and between schools in teacher RISE, 7Cs, and VAM ratings

		Percentage of variance[a]		
Measure	Number of schools	Within schools	Between schools	Number of teachers
Composite of 24 RISE components	63	84.9	15.1	1,060
Composite 7Cs rating	60	95.3	4.7	1,684
VAM reading	51	88.6	11.4	170
VAM math	49	91.1	8.9	139

RISE is Research-based Inclusive System of Evaluation. VAM is value-added measure.

Note: Science and social studies VAMs are not included because of the small number of teachers with VAM estimates in those subjects.

a. Calculated using analysis of variance weighted by the number of teachers in each school.

Source: RISE and 7Cs data are from Pittsburgh Public Schools for the 2011/12 school year; VAM data are from Johnson et al. (2012) and are for 2008–11.

correlates strongly (at 0.96 and above) with the alternatives considered (a 24-component average, an average of the Danielson Framework for Teaching components, and an average of the 12 power components weighted, based on the principal components analysis).

Teachers of low-income and racial/ethnic minority students tend to receive lower RISE and 7Cs ratings, and teachers of gifted students tend to receive higher RISE and 7Cs ratings

By student subgroup, RISE and 7Cs ratings are negatively correlated with the percentage of students eligible for free or reduced-price lunch and the percentage of racial/ethnic minority students and positively correlated with the percentage of students designated as gifted. RISE and 7Cs composites are not related to the percentages of English language learner students, students in special education, or female students (table 3).[3] The significant correlations could indicate that teacher quality is inequitably distributed across classrooms or that student characteristics directly affect teacher ratings, making the ratings biased measures of teacher quality. Several years of RISE ratings for each teacher are required to be able to determine which explanation applies. Such data will not be available until the end of the 2013/14 school year.

Table 3. Correlations of RISE and 7Cs ratings with percentages of students in various subgroups

Measure	Racial/ ethnic minority	Free or reduced-price lunch	English language learner	Special education	Gifted	Female
Composite of 12 RISE power components (932 teachers)	**–.20**	**–.15**	.05	.05	**.12**	–.02
Composite 7Cs Score (1,535 teachers)	**–.14**	**–.09**	.05	.03	**.09**	–.03

RISE is Research-based Inclusive System of Evaluation.

Note: These are simple Pearson correlations. Bold denotes statistically significant at the .05 level.

Source: Pittsburgh Public Schools data for the 2011/12 school year.

Similarly, future student-level 7Cs data could be used to explore the extent to which individual teachers receive different scores from different types of students to help control for the independent effect of student characteristics on those measures. Any remaining variation in the 7Cs ratings across teachers could then more easily be attributed to the teacher rather than to the students.

Few of the correlations of VAM estimates with student characteristics are statistically significant (table B12 in appendix B). The few significant correlations are consistent with those reported for RISE and 7Cs: racial/ethnic minority students, students from a low-income household, and nongifted students are often served by teachers with lower than average VAM estimates. Since VAMs control for the characteristics of individual students, the correlations suggest some inequality in the distribution of teachers across students rather than bias in the measure.

Higher RISE ratings are associated with some teacher characteristics

All else equal, English teachers and elementary school teachers earn higher RISE ratings than other teachers do (table 4). Special education teachers also perform slightly better than other teachers. It is difficult to know whether these differences reflect true differences in average teacher effectiveness or merely differences in the standards used to evaluate different types of teachers.

All else equal, English teachers and elementary school teachers earn higher RISE ratings than other teachers do, and teachers with more experience receive higher RISE ratings than do teachers with less experience

Teachers with more experience receive higher RISE ratings than do teachers with less experience (see table 4). This finding is consistent with findings from prior research suggesting that teacher effectiveness improves over the first few years of teaching. The positive relationship between experience and measured effectiveness is also consistent with VAM

Table 4. Coefficient estimates for RISE rating regressed on teachers by subject taught and years of experience

							Years of teacher experience			
RISE measure	Math	English	Science	Foreign language	Elementary teachers of multiple subjects	Special education	2–3	4–8	9–20	More than 20
Planning and preparation domain	.05	**.12**	.06	.05	.05	.04	**.10**	**.17**	**.17**	**.23**
Classroom environment domain	.04	**.17**	.08	.00	**.15**	.09	.03	**.14**	**.12**	**.15**
Teaching and learning domain	.05	**.14**	.04	.09	**.09**	.07	**.09**	**.14**	**.15**	**.22**
Professional responsibilities domain	.02	.08	.08	.02	**.09**	.06	.06	.05	.07	**.09**
12 power components	.06	**.14**	.06	.06	**.12**	**.10**	.08	**.14**	**.15**	**.20**

RISE is Research-based Inclusive System of Evaluation.

Note: Bold denotes statistically significant at the .05 level. For years of teacher experience, the comparison group is teachers with less than two years of experience. Teachers of each subject are compared with those of all other subjects. The comparison for elementary school teachers of multiple subjects includes all departmentalized teachers across grades. For each regression the sample is 1,014 teachers, and the *r*-squared statistic varies from .049 to .061, except for the professional responsibilities domain, which has one fewer teacher (1,013) and an *r*-squared of .030.

Source: Pittsburgh Public Schools data for the 2011/12 school year.

results (table B13 in appendix B).[4] Contrary to what is typically found in other districts, however, the RISE and VAM results for Pittsburgh show continued increases in effectiveness even at very high levels of experience: teachers with more than 20 years of experience generally have the highest RISE and VAM ratings.

There are at least three possible explanations for this pattern. First, Pittsburgh's teachers might be continually improving their practice over time, even though continuous improvement has not been evident in other districts and states. Second, highly effective teachers might be more likely to continue teaching in Pittsburgh Public Schools, regardless of whether their performance improves over time. The strong results for the most experienced teachers could result from a steady attrition of less effective teachers. Third, the district might have been able to recruit more-effective teachers 20 years ago than it can today. If the effectiveness of newly hired teachers in the district has declined steadily, then the most experienced group would be most effective. Which of these is the correct explanation (or combination of explanations) cannot be determined without collecting and analyzing enormous amounts of very long-term data.

Despite the clear relationship between teacher experience and VAM estimates, experience generally explains less than 10 percent of the variation in VAM estimates (see table B13). Thus, even though more-experienced teachers have higher average effectiveness, effective and ineffective teachers can be found at all levels of experience.

Correlations between the RISE and 7Cs component and composite ratings are all positive and usually statistically significant, though modest in size

Relationships among teacher evaluation ratings

This section examines the relationships between and among measures of professional practice (RISE rating), student surveys (7Cs rating), and student achievement growth (VAM estimates) using methods similar to those of other analyses for Pennsylvania (Lipscomb et al., 2012) and elsewhere (Kimball, White, Milanowski, & Borman, 2004; Kane & Staiger, 2012). Positive relationships would provide cross-validation of the measures. These analyses answer the last three research questions (5–7).

Principals who were preparing RISE ratings in 2011/12 could not have seen 7Cs student survey results or VAM estimates for individual teachers because all three measures were under development or in pilot phases at the time. Principals had prior access to schoolwide VAM estimates by grade and subject, but this would not help in evaluating individual teachers, because most of the variation in teacher value added is within schools.

RISE and 7Cs ratings are positively correlated

Correlations between the RISE and 7Cs component and composite ratings are all positive and usually statistically significant, though modest in size (table 5). Indeed, the RISE composite rating for the 12 power components is significantly correlated with every one of the seven 7Cs component ratings, and the 7Cs composite rating is significantly correlated with every one of the 24 RISE component ratings. The RISE component with the weakest relationship with 7Cs ratings is component 4e, "growing and developing professionally": only one of the correlations with the components of the 7Cs is statistically significant, perhaps suggesting the need to scrutinize this component of RISE—unless the district does not expect professional growth to be correlated with student perceptions of teachers' classroom practices.

Table 5. Correlations between RISE and 7Cs rating

		7Cs measure							
RISE domain	**RISE component**	**Cares**	**Controls**	**Clarifies**	**Challenges**	**Captivates**	**Confers**	**Consolidates**	**Average**
1: Planning and preparation	1a. Demonstrating knowledge of content and pedagogy	.06	**.07**	**.08**	**.10**	.06	**.08**	.06	**.08**
	1b. Demonstrating knowledge of students*	**.20**	**.16**	**.18**	**.19**	**.19**	**.17**	**.15**	**.20**
	1c. Setting instructional outcomes*	**.22**	**.19**	**.20**	**.18**	**.21**	**.18**	**.16**	**.22**
	1d. Demonstrating knowledge of resources	**.09**	**.12**	**.08**	**.10**	**.10**	**.08**	.04	**.10**
	1e. Planning coherent instruction	**.08**	**.09**	**.14**	**.12**	**.08**	**.09**	**.10**	**.12**
	1f. Designing ongoing formative assessments	**.09**	**.12**	**.14**	**.10**	.05	**.11**	**.11**	**.12**
2: Classroom environment	2a. Creating learning environment of respect and rapport	**.25**	**.29**	**.24**	**.23**	**.27**	**.23**	**.23**	**.29**
	2b. Establishing a culture for learning*	**.23**	**.23**	**.25**	**.23**	**.21**	**.22**	**.21**	**.26**
	2c. Managing classroom procedures	**.16**	**.19**	**.18**	**.18**	**.15**	**.11**	**.15**	**.19**
	2d. Managing student behavior*	**.29**	**.32**	**.31**	**.30**	**.29**	**.28**	**.28**	**.34**
	2e. Organizing physical space	**.10**	**.09**	**.09**	**.10**	**.11**	.07	**.08**	**.11**
3: Teaching and learning	3a. Communicating with students	**.18**	**.18**	**.15**	**.15**	**.16**	**.18**	**.14**	**.19**
	3b. Using questioning and discussion techniques*	**.11**	**.14**	**.14**	**.14**	**.09**	**.09**	**.11**	**.14**
	3c. Engaging students in learning*	**.19**	**.20**	**.19**	**.15**	**.18**	**.14**	**.16**	**.20**
	3d. Using assessment to inform instruction*	**.14**	**.14**	**.17**	**.12**	**.09**	**.15**	**.16**	**.16**
	3e. Demonstrating flexibility and responsiveness	**.13**	**.10**	**.14**	**.11**	**.12**	**.14**	**.11**	**.14**
	3f. Assessment results and student learning*	**.15**	**.14**	**.17**	**.17**	**.12**	**.13**	**.14**	**.17**
	3g. Implementing lessons equitably*	**.16**	**.15**	**.13**	**.11**	**.14**	**.10**	**.11**	**.15**

(continued)

Table 5. Correlations between RISE and 7Cs rating *(continued)*

		7Cs measure							
RISE domain	RISE component	Cares	Controls	Clarifies	Challenges	Captivates	Confers	Consolidates	Average
4: Professional responsibilities	4a. Reflecting on teacher and student learning*	**.13**	**.12**	**.12**	**.13**	**.10**	**.13**	**.10**	**.14**
	4b. Systems for managing student data*	**.15**	**.16**	**.16**	**.15**	**.10**	**.11**	**.13**	**.16**
	4c. Communicating with families*	**.17**	**.12**	**.17**	**.19**	**.14**	**.16**	**.15**	**.18**
	4d. Participating in a professional community	.06	.06	**.09**	**.07**	**.09**	**.10**	**.09**	**.09**
	4e. Growing and developing professionally	.05	.07	.05	.06	**.09**	.06	.06	**.07**
	4f. Showing professionalism	**.12**	**.12**	**.09**	**.13**	**.11**	**.09**	.06	**.12**
Composite	12 power components	**.27**	**.27**	**.29**	**.27**	**.24**	**.24**	**.24**	**.30**

RISE is Research-based Inclusive System of Evaluation.

* One of the 12 power components.

Note: 7Cs measures are based on student surveys. Sample size ranges from 754 to 894 teachers. Bold denotes statistically significant at the .05 level.

Source: Pittsburgh Public Schools data for the 2011/12 school year.

RISE ratings and VAM estimates are positively correlated

Correlating RISE ratings and 7Cs ratings with VAM estimates in the same year could overestimate true long-term correlations, because unusual circumstances (for example, a particularly disruptive student) could result in errors that point in the same direction for all three measures. Such a possible correlation of errors across measures in the same classroom for a given year is avoided by including data for multiple years. The analyses use an average of up to three years of value-added data covering 2008–11; RISE and 7Cs data are for the 2011/12 academic year.

The correlations between RISE component and composite ratings and VAM estimates are nearly always positive—with the notable exception of VAM estimates for social studies courses using curriculum-based assessments, where most of the correlations are close to zero and none is statistically significant (table 6). The positive correlations are not large and often are not statistically significant, but their consistency is clear. The statistical power to detect relationships is lower than it is for results based on correlations of RISE with 7Cs, because VAM estimates are not available for most teachers (because they do not teach grades and subjects for which VAMs can be estimated). When VAM estimates are aggregated across all grades and subjects (the VAM "all grade levels" column in table 6), the correlations of the RISE ratings with the VAM estimates range from .02 to .23, and most (15 out of 25) are statistically significant. The relationships are statistically significant most consistently in English language arts and for VAM estimates based on the Pennsylvania System of School Assessment.

Table 6. Correlations between RISE and prior-year value-added measure estimates

		Value-added measure									
RISE domain	RISE component	English language arts	Math	Science	Social studies	All grade levels	Elementary school	Middle school	High school	Pennsylvania System of School Assessment	Curriculum-based assessment
1: Planning and preparation	1a. Demonstrating knowledge of content and pedagogy	.16	.13	.07	–.18	**.12**	.20	.11	–.01	**.15**	.03
	1b. Demonstrating knowledge of students*	.13	.04	.20	–.03	.07	.04	.10	.10	.08	.11
	1c. Setting instructional outcomes*	**.17**	.08	.19	–.10	**.14**	.17	.11	.03	**.17**	.07
	1d. Demonstrating knowledge of resources	**.17**	–.03	.10	–.25	.10	.10	.13	–.00	.12	.04
	1e. Planning coherent instruction	**.18**	**.19**	.09	–.08	**.11**	.15	.12	.14	**.14**	.13
	1f. Designing ongoing formative assessments	.15	.13	.13	–.09	.08	.10	**.16**	–.04	.11	.06
2: Classroom environment	2a. Creating learning environment of respect & rapport	**.23**	**.21**	.15	.04	**.16**	.15	**.23**	–.00	**.22**	**.14**
	2b. Establishing a culture for learning*	**.27**	**.23**	.18	.09	**.23**	**.25**	**.19**	.16	**.29**	.10
	2c. Managing classroom procedures	.15	**.24**	**.26**	.09	**.18**	**.21**	.11	.13	**.16**	.08
	2d. Managing student behavior*	**.23**	**.17**	**.22**	.07	**.21**	**.22**	.12	**.23**	**.19**	.12
	2e. Organizing physical space	.10	.03	.12	–.07	.08	.08	.02	.16	.06	.03
3: Teaching and learning	3a. Communicating with students	.15	.16	.20	–.01	**.12**	.09	.15	.07	**.17**	**.14**
	3b. Using questioning and discussion techniques*	**.18**	**.17**	**.25**	.09	**.18**	**.22**	**.20**	.03	**.20**	**.13**
	3c. Engaging students in learning*	**.20**	**.18**	**.32**	–.14	**.21**	**.35**	.11	.04	**.28**	.01
	3d. Using assessment to inform instruction*	**.18**	.12	–.00	–.13	.08	.11	.06	.03	.08	.06
	3e. Demonstrating flexibility and responsiveness	.15	**.21**	**.32**	.07	**.17**	**.23**	**.16**	.07	.12	**.20**
	3f. Assessment results and student learning*	**.20**	.05	**.31**	.08	**.17**	.16	.13	.16	**.16**	**.15**
	3g. Implementing lessons equitably*	.13	.16	.20	–.03	**.12**	.14	.11	.05	**.14**	-.01
4: Professional responsibilities	4a. Reflecting on teacher and student learning*	**.23**	–.01	–.01	–.15	.05	.07	.13	–.09	.08	.07
	4b. Systems for managing student data*	**.15**	.06	.05	–.22	.06	.07	.05	.04	.06	.06
	4c. Communicating with families*	.12	–.00	.07	–.01	.07	.01	.11	.03	.06	.06
	4d. Participating in a professional community	**.19**	.03	.15	–.12	**.12**	.11	.15	.11	**.14**	**.13**
	4e. Growing and developing professionally	.13	–.08	.08	–.16	.02	.04	.05	.02	.06	.02
	4f. Showing professionalism	.05	.05	.18	–.17	.08	.13	.07	–.06	**.13**	.02
Composite	12 power components	**.29**	.16	**.27**	–.06	**.22**	**.24**	**.19**	.11	**.24**	**.13**

RISE is Research-based Inclusive System of Evaluation.

* One of the 12 power components.

Note: Bold denotes statistically significant at the .05 level. Sample sizes range from 41 to 358 teachers and vary in ways similar to those shown in table D3 in appendix D.

Source: RISE data are from Pittsburgh Public Schools for the 2011/12 school year; Value-added data are from Johnson et al. (2012) and are for 2008–11.

Not surprisingly, the RISE composite rating based on the 12 component average more often shows statistically significant relationships with VAM estimates than do most of the individual RISE component ratings. Indeed, the RISE composite rating is significantly correlated with the VAM estimate in 70 percent of the VAM groups. When the analysis for the VAM estimate is done across all grades and subjects (the VAM "all grade levels" column in table 6), the RISE composite correlation with the VAM estimate is .22.

The correlations found here are larger than those found in a study by the Measures of Effective Teaching project (Bill & Melinda Gates Foundation, 2010), which conducted its analyses by subject. It found that a composite Framework for Teaching measure correlated with "underlying" value added (after adjusting the correlations upward to account for estimation error) at .11 in reading and .18 in math (Kane & Staiger, 2012). The correlations calculated in the current study are .29 in reading and .15 in math for the Danielson-based composite rating (not reported in table 6) without adjusting upward for estimation error. The correlations found for Pittsburgh Public Schools are comparable to those found by Milanowski (2011) using data for school districts in Cincinnati, OH; Coventry, RI; and Washoe County, NV.

The two RISE components with the largest number of statistically significant relationships with the VAM estimates are managing student behavior and using questioning and discussion techniques

For individual RISE component ratings, correlations with VAM estimates vary considerably. Five of the 24 components, including 2 of 12 power components, have no statistically significant associations with VAM estimates in any subject or grade level. Three of the components showing no statistically significant correlation with VAM estimates are in domain 4, professional responsibilities. The two RISE components with the largest number of statistically significant relationships with the VAM estimates are 2d, managing student behavior, and 3b, using questioning and discussion techniques. Both are power components.

Across grades and subjects, ratings for component 3f, assessment results and student learning, have a correlation of .17 with the all-grade-levels VAM rating—higher than the average correlation of .10 across the 24 RISE components.[5] Only 5 of the 24 components—all in domains 2 and 3—have higher correlations with the VAM estimates across grades and subject.

The ratings for component 4e, growing and developing professionally, which are only weakly correlated with 7Cs ratings, are also weakly correlated with VAM estimates, and none of the correlations is statistically significant. The average correlation for RISE component 4e with VAM estimates (.02) is lower than for any of the other RISE components.

7Cs and value-added measure estimates are positively correlated

The correlations between the 7Cs rating and the VAM estimates are weaker than the correlations between RISE and 7Cs rating, but the sample sizes are also smaller (table 7). Overall, 56 percent of the correlations are statistically significant compared with more 90 percent of the correlations between RISE and 7Cs ratings. All of the significant correlations and most of the nonsignificant correlations are positive.

Some clear patterns emerge in the correlations between the 7Cs rating and the VAM estimates. First, control and challenge components of the 7Cs have more statistically significant correlations with VAM estimates than do the other components. This suggests that

Table 7. Correlations between 7Cs rating and prior-year value-added measure estimates

	Value-added measure									
7Cs domain	English language arts	Math	Science	Social studies	All grade levels	Elementary school	Middle school	High school	Pennsylvania System of School Assessment	Curriculum-based assessment
Cares	.11	.10	.02	.07	**.09**	–.02	.09	**.15**	.06	**.13**
Controls	**.20**	**.14**	**.19**	.15	**.17**	.09	**.18**	**.23**	**.16**	**.20**
Clarifies	**.18**	.10	.07	.08	**.13**	.04	**.13**	**.16**	**.12**	**.14**
Challenges	**.25**	.10	.03	.14	**.18**	**.18**	**.13**	**.18**	**.19**	**.11**
Captivates	.10	.08	.15	.10	**.12**	–.05	**.13**	**.21**	.08	**.17**
Confers	**.16**	.06	.01	.13	**.11**	.00	.11	**.22**	**.11**	**.14**
Consolidates	**.13**	.11	.05	.15	**.13**	.01	**.14**	**.18**	**.12**	**.12**
7Cs composite[a]	**.18**	.11	.09	.13	**.15**	.04	**.14**	**.21**	**.14**	**.16**

Note: 7Cs measures are based on student survey data. Bold denotes statistically significant at the .05 level. Sample sizes range from 59 teachers for VAM social studies to 619 teachers for VAM all grade levels and vary in ways similar to those shown in table D4 in appendix D.

a. Average across all seven components.

Source: 7Cs data are from Pittsburgh Public Schools for the 2011/12 school year; Value-added data are from Johnson et al. (2012) and are for 2008–11.

control and challenge measures might be particularly important to examine for teachers lacking VAM estimates.

Second, the relationships between RISE ratings and VAM estimates differ in important ways from the relationships between 7Cs ratings and VAM estimates, suggesting that the RISE and 7Cs measures provide complementary information about teacher skills. For RISE ratings and VAM estimates, the correlations are somewhat stronger for elementary schools than for middle and high schools and for the Pennsylvania System of School Assessment than for the locally developed, curriculum-based assessments. For 7Cs ratings and VAM estimates, in contrast, correlations are generally positive and statistically significant for middle and high school but not for elementary school, and correlations are generally stronger for VAM estimates that rely on curriculum-based assessments than for estimates that rely on the state assessment.

VAM estimates account for student characteristics, while RISE and 7Cs ratings do not. The lack of controls for student characteristics might affect the correlations reported here, particularly as the analysis found statistically significant relationships between student characteristics and RISE and 7Cs ratings (see previous section).[6] Analyzing several years of RISE and 7Cs data could establish more conclusively whether the lack of adequate controls for student characteristics at the classroom level might weaken the relationships between RISE ratings and VAM estimates. The availability of student-level 7Cs data might enable a similar analysis for 7Cs using only one year of data.

Because true teacher effectiveness probably changes over time, calculating correlations of measures from different school years could underestimate the true correlations of RISE and 7Cs ratings with VAM estimates. This possibility was tested by calculating separate correlations of RISE and 7Cs rating with VAM estimates that include the 2011/12 school year. The results (available in appendix D) do not differ in substantively important ways from the results presented in tables 6 and 7, which do not include data for that school year.

RISE ratings suggest partial but imperfect consistency of standards among principals

A correlational analysis of RISE ratings with VAM estimates helps assess whether differences in principals' ratings reflect true differences in teacher effectiveness or inconsistencies in principals' calibration of rating standards. For example, some principals might assign their teachers higher (or lower) RISE ratings than the ratings the teachers receive on VAM and 7Cs, suggesting that those principals may be producing RISE ratings that are biased upward (or downward). Other principals might assign RISE ratings that are uncorrelated with VAM estimates and 7Cs ratings, suggesting that those principals might be inconsistent in applying RISE ratings. The first pattern would raise concerns about the validity of RISE ratings and the second would raise concerns about the reliability of RISE ratings. Either pattern might suggest the need for additional training for principals in the application of RISE ratings.

As noted earlier, most of the variation in RISE ratings and VAM estimates occurs within schools rather than between them, a similarity that would be expected if principals were rating teachers correctly. Within-school correlations by school between RISE ratings and VAM estimates could also shed light on the consistency of principals' ratings, but there are too few teachers per school who have both RISE ratings and VAM estimates. Those data can, however, be used to look at the relationships within schools of RISE ratings with 7Cs ratings and VAM estimates by combining results across all schools. Analyzed in this way, the data show evidence of consistency of scaling across schools. But there is also evidence that the rater matters; the scale consistency across schools/raters is not perfect. Principals' RISE ratings of teachers partially align with the 7Cs ratings and VAM estimates within and across schools, but the RISE ratings also vary systematically across schools in ways that are not explained by the 7Cs ratings and VAM estimates (as described in appendix D). This could be the result of true school-level differences in average teacher performance that are not captured by RISE and 7Cs ratings, but a simpler explanation is that some of Pittsburgh's principals have higher or lower standards than the others.

The data show evidence of consistency of scaling across schools, but there is also evidence that the rater matters; the scale consistency across schools/raters is not perfect

Study implications, limitations, and suggestions

Despite some limitations (see below), this study makes a valuable contribution to the research on relationships among multiple measures of teacher performance, which is still quite new. Evidence of significantly positive correlations among the three measures reinforces previous findings (such as those from the Gates Foundation's Measures of Effective Teaching project), confirming that those findings remain relevant in other contexts.

The findings in this report demonstrate that with training and a system for gathering rich evidence on professional practice, single-rater systems can produce ratings of professional practice that predict teacher value added. The fact that RISE measures are significantly correlated with estimates of teachers' value added, despite relying on only a single rater, provides encouraging evidence of the rigor of the RISE process. The findings of the Measures of Effective Teaching project suggested that it might not be possible to achieve reliable measures without using several raters (Kane & Staiger, 2012). In Pittsburgh there is evidence that a single rater can produce useful measures in a context that includes an extensive process of evidence collection throughout the year, giving the rater much more information on teachers than an anonymous reviewer of a classroom video would have (as was the case in the Measures of Effective Teaching project).

Study limitations

The fact that the correlations of RISE and 7Cs ratings with VAM estimates are only moderate suggests that the RISE and 7Cs ratings may capture teacher skills that affect student outcomes that are not measured in the tests used for the VAM estimates. For example, RISE and 7Cs ratings might capture teachers' contributions to students' creativity, perseverance, and positive behaviors, which may not be perfectly correlated with test scores. Hence, the fact that some of the RISE and 7Cs components are not significantly correlated with VAM estimates does not necessarily mean that the components are not valid.

When RISE ratings are used for high-stakes decisions, the validity of the ratings may decline if teachers focus on superficial ways of increasing their RISE ratings without truly improving their performance.

The fact that the correlations of RISE and 7Cs ratings with VAM estimates are only moderate suggests that the RISE and 7Cs ratings may capture teacher skills that affect student outcomes that are not measured in the tests used for the VAM estimates

Some of the findings of this study may be unique to the particular circumstances and measures used in Pittsburgh Public Schools. There is no way to know, for example, whether correlations between RISE and 7Cs ratings and VAM estimates based on Pittsburgh's local curriculum-based assessments will apply in other districts using different assessments. Correlations could also vary with the number of years of teaching covered by the VAM estimates and according to whether multiyear estimates are weighted equally (as Pittsburgh has chosen to do) or weighted more heavily toward the current year. And finally, the training and practices of raters (principals or others) could substantially affect the results.

A final limitation of the study is that nearly half of teachers did not have data on the 12 power components of RISE (tables B1, B2) and that those missing some data appear to differ from those with more complete data (table B14). Each year a third of teachers participate in supported growth plans that involve a deep focus on one RISE component and therefore they do not receive ratings on the other components. Meanwhile, teachers whose performance was found to be unsatisfactory in the preceding year and who were put on employee improvement programs were not given RISE ratings in 2011/12. (This has since changed, and teachers on employee improvement plans now receive RISE ratings.) Excluding these teachers likely reduces the correlation between RISE ratings and VAM estimates by eliminating a group that is likely to be in the lower tail of the distribution on both measures.

Suggestions for improving the system of multiple measures of teacher effectiveness

RISE could be improved by using multiple raters for each teacher. Even though Pittsburgh's RISE system already produces a measure that predicts estimates of teacher value added, there is room for improvement. There is some evidence from variations between schools in the correlations between RISE ratings and VAM estimates that principals are not entirely consistent in how they apply the RISE rubric. Using additional raters to work with principals in evaluating each teacher would improve calibration of the RISE scale across schools, particularly if some of the additional raters work in multiple schools. Using multiple raters would also reduce the potential influence of irrelevant subjective factors. Moreover, the Measures of Effective Teaching project demonstrates that using multiple raters can substantially increase the reliability of the ratings (Kane & Staiger, 2012).

Future analyses could contribute to the continuous improvement of Pittsburgh Public Schools' system of multiple measures of teacher effectiveness. The study found significant correlations of the RISE and 7Cs ratings with student characteristics, suggesting the desirability of further analyses to determine whether the RISE and 7Cs measures should be adjusted for student characteristics, much as VAM estimates are. The currently available data do not allow a definitive recommendation because they cannot distinguish between correlations caused by true differences in teacher quality and correlations created by bias (a result of something about the students that is outside a teacher's control). Distinguishing between those two explanations requires additional data to assess whether the same teachers earn different ratings when they serve different kinds of students. Analyzing the student-level data would permit the necessary analysis for 7Cs. Because RISE ratings are not produced from student-level data, it would be necessary to include additional years of data to assess whether individual teachers' RISE ratings vary with the classroom averages of the characteristics of the students they serve.

Additional data are required to assess whether the same teachers earn different ratings when they serve different kinds of students

Additional years of data would also permit further analysis using the different measures to predict student outcomes. This information could be used to develop better composite measures combining all three sources of information and to assess the extent to which each measure independently predicts future growth in student achievement.

Appendix A. Descriptions of the three teacher evaluation measures

This appendix describes the three measures used for teacher evaluation in Pittsburgh Public Schools.

The Research-based Inclusive System of Evaluation measure of professional practice

The Research-based Inclusive System of Evaluation (RISE) is an observational measure of teacher practice, based on ratings by the school principal. Data for 2011/12 include teachers' ratings on components that are divided into four domains and 24 components (table A1):

- *Planning and preparation:* Demonstrating knowledge and planning for instruction and assessments.
- *Classroom environment:* Managing the learning environment and behavior of students.
- *Teaching and learning:* Communicating with and engaging students while promoting equity, and producing achievement growth.
- *Professional responsibilities:* Relating to work outside the classroom, including communicating with families and colleagues and growing professionally.

Table A1. RISE components by domain

Domain	Component
1: Planning and preparation	1a. Demonstrating knowledge of content and pedagogy
	1b. Demonstrating knowledge of students*
	1c. Setting instructional outcomes*
	1d. Demonstrating knowledge of resources
	1e. Planning coherent instruction
	1f. Designing ongoing formative assessments
2: Classroom environment	2a. Creating a learning environment of respect & rapport
	2b. Establishing a culture for learning*
	2c. Managing classroom procedures
	2d. Managing student behavior*
	2e. Organizing physical space
3: Teaching and learning	3a. Communicating with students
	3b. Using questioning and discussion techniques*
	3c. Engaging students in learning*
	3d. Using assessment to inform instruction*
	3e. Demonstrating flexibility and responsiveness
	3f. Assessment results and student learning*
	3g. Implementing lessons equitably*
4: Professional responsibilities	4a. Reflecting on teacher and student learning*
	4b. Systems for managing student data*
	4c. Communicating with families*
	4d. Participating in a professional community
	4e. Growing and developing professionally
	4f. Showing professionalism

RISE is Research-based Inclusive System of Evaluation.

* One of the 12 power components.

Source: Pittsburgh Public Schools.

The RISE components were developed based on Charlotte Danielson's Framework for Teaching (Danielson, 2013) in a collaborative process between district teachers and administrators. RISE includes two additional components that are not part of the Danielson framework (3f, assessment results and student learning; 3g, implementing lessons equitably) as well as slight modifications to some of the framework's other components.

Of the 24 components, the district identified 12 as "power components" for the 2011/12 school year (see table A1). During the RISE design phase, principals and lead teachers in pilot schools reported that they believed that these 12 components have the greatest impact on student learning and growth. Although all 24 components were acknowledged to be important to teaching, the rubric was reduced to 12 to lighten the burden of evaluation on principals and teachers alike. Most teachers were evaluated on the 12 power components; a smaller fraction was evaluated on all 24. The specific power components may change in future years.

Principals categorize teachers' performance on each RISE component as unsatisfactory, basic, proficient, or distinguished. Principals also give most teachers an overall RISE rating that is based on the principal's judgment rather than on an average of the component ratings. The ratings are converted to numerical scores of 0, 1, 2, and 3 for analytic purposes. This report also considers various composite measures of the component ratings, such as an unweighted average of the power component scores, an unweighted average of all RISE component ratings, and in a few cases weighted averages of the component ratings. In most cases the primary composite measure analyzed for this report is an equally weighted (unweighted) average of the 12 power components.

Each school principal assigns RISE ratings to all teachers in the school, based on formal and informal observations conducted at least four times a year. RISE ratings are also informed by evidence of teaching practice collected by teachers throughout the school year and by teacher self-evaluations conducted at the beginning, middle, and end of the school year. At the end of each year principals create summative ratings for each teacher based on all the information obtained by then.

Principals were the sole RISE raters in their schools during the 2011/12 school year and did not rate teachers in other schools.[7] The district sought to promote consistency in ratings by specifying standards for each rating level on each component and by providing extensive training for principals and teachers. As part of the district's Instructional Quality Assurance and Certification process, all principals received at least two years of training on RISE standards, including five video observations as a pre-assessment, three to five days of online training in the fundamentals of the Danielson Proficiency System, up to five days of coaching, an additional five video observations as a post-assessment, and up to five more days of coaching. In addition, a few days were focused on the certification process. Principals were also trained to provide constructive feedback to improve teaching effectiveness. At the end of training principals had to demonstrate that they could assign accurate and consistent ratings. After the initial two years of training, principals participate in ongoing recertification, focusing on a sample of teachers and observing actual changes in practice over the course of a year.

7Cs student survey measure

The 7Cs surveys were developed by Ronald Ferguson of Harvard University as part of the Tripod Project and delivered by Cambridge Education (Kane & Staiger, 2012). Students assess teachers by responding to survey questions organized under seven constructs based on the Tripod 7Cs framework:

- *Caring for students:* Being encouraging and supporting.
- *Controlling the classroom:* Maintaining a culture of cooperation and peer support.
- *Clarifying messages:* Helping students see that success is feasible.
- *Challenging learners:* Pressing students to exert effort, persevere, and be rigorous.
- *Captivating the classroom:* Making learning interesting and relevant.
- *Conferring with students:* Showing respect for students' ideas.
- *Consolidating lessons learned:* Connecting and integrating course contents.

Students indicate whether they agree or disagree with statements about their teacher's ability to perform tasks related to each of the 7Cs. Each 7C component is measured using responses to the survey items for that component, with the ratings reported in normal-curve equivalents among teachers in Pittsburgh Public Schools. The 7Cs surveys are administered using different surveys for each grade band (K–2, 3–5, 6–8, and 9–12), so teacher averages were normalized within grade bands to generate the normal-curve equivalents.[8] An average of the 7Cs component ratings for each teacher is also reported. Normal-curve-equivalent ratings are equivalent to percentiles near the middle (50) and ends (1 and 99) of the distribution. Unlike percentiles, however, normal-curve equivalents are on an equal-interval scale: for example, the difference between 50 and 60 is the same as the difference between 80 and 90. And unlike percentiles, normal-curve equivalents can be meaningfully averaged.

Other research has shown the 7Cs measures to be reliable and stable for the same teachers across several administrations of the surveys in the same year and highly correlated across different class sections taught by the same teacher (Kane, 2012; Kane & Staiger, 2012; Bill & Melinda Gates Foundation, 2010). The same studies have also found the 7Cs measures to be correlated with estimates of teacher value added in locations outside Pennsylvania and using different student assessments.

Value-added measures

This report employs the teacher value-added measure (VAMs) estimates calculated by Johnson et al. (2012) using Pittsburgh Public Schools' longitudinal student-level data. VAMs were calculated by grade (4–12), subject (math, reading/English language arts, social studies, and science), and test type (Pennsylvania System of School Assessment and curriculum-based assessments) for 2008–11 (table A2). In addition to controlling for previous achievement, the VAMs adjust for student background characteristics by incorporating data on poverty, special needs, gifted status, race/ethnicity, grade repetition, and English language learner status. VAM estimates are reported in normal-curve-equivalent units.

Table A2. Assessment outcome and grade used in Pittsburgh Public Schools value-added measures

Elementary grades	Middle grades	High school grades
PSSA Math Grade 3	PSSA Math Grade 6	CBA Algebra I/AB-BC Grade 9
PSSA Reading Grade 3	PSSA Reading Grade 6	CBA ELA I Grade 9
PSSA Math Grade 4	CBA Math Grade 6	CBA Biology Grade 9
PSSA Reading Grade 4	CBA English Grade 6	CBA Civics Grade 9
PSSA Science Grade 4	CBA Earth Science Grade 6	CBA Geometry/AB-BC Grade 10
PSSA Math Grade 5	PSSA Math Grade 7	CBA ELA II Grade 10
PSSA Reading Grade 5	PSSA Reading Grade 7	CBA Chemistry Grade 10
PSSA Writing Grade 5	CBA Math Grade 7	CBA World History Grade 10
	CBA English Grade 7	CBA Algebra 2 Grade 11
	CBA Life Science Grade 7	CBA ELA III Grade 11
	PSSA Math Grade 8	CBA Physics Grade 11
	PSSA Reading Grade 8	CBA US History Grade 11
	PSSA Science Grade 8	CBA ELA IV/AA Lit Grade 12
	PSSA Writing Grade 8	
	CBA Math Grade 8	
	CBA English Grade 8	
	CBA Physics Grade 8	
	CBA US History Grade 8	

PSSA is Pennsylvania System of School Assessment. CBA is curriculum-based assessment. ELA is English language arts.

Source: Data are from Johnson et al. (2012) and are for 2008–11.

Appendix B. Detailed tables

Table B1. Number of teachers included in the analyses

Group	Number of teachers	Description
1	2,082	Teachers in regular schools
2	1,625	Teachers with any RISE data
3	1,975	Subset of 1 in schools that did not close
4	1,069	Subset of 2 with at least 12 RISE components
5	358	Subset of 4 with 2008–11 value-added data
6	894	Subset of 4 with 7Cs data
7	329	Subset of 4 with value-added and 7Cs data

RISE is Research-based Inclusive System of Evaluation.

Source: RISE and 7Cs data are from Pittsburgh Public Schools for the 2011/12 school year; Value-added data are from Johnson et al. (2012) and are for 2008–11.

Table B2. Number and percentage of teachers in regular schools, by number of RISE scores

Number of RISE scores	Number of teachers	Percentage of teachers	Cumulative
0	457	22.0	22.0
1	554	26.6	48.6
2	2	0.1	48.7
12	1	0.1	48.7
13	90	4.3	53.0
14	23	1.1	54.1
15	22	1.1	55.2
16	13	0.6	55.8
17	5	0.2	56.1
18	1	0.1	56.1
19	1	0.1	56.2
20	4	0.2	56.3
21	2	0.1	56.4
22	5	0.2	56.7
23	20	1.0	57.6
24	80	3.8	61.5
25	802	38.5	100.0
Total	2,082	100.0	100.0

RISE is Research-based Inclusive System of Evaluation.

Note: Values may not sum to totals because of rounding.

Source: Data are from Pittsburgh Public Schools for the 2011/12 school year.

Table B3. Percentage of teachers, by RISE rating

		RISE rating[a]				Number of teachers with component
Domain	Component	0	1	2	3	
1: Planning and preparation	1a. Demonstrating knowledge of content and pedagogy	0	7	82	11	917
	1b. Demonstrating knowledge of students*	0	7	78	16	1,069
	1c. Setting instructional outcomes*	0	8	83	9	1,069
	1d. Demonstrating knowledge of resources	0	7	85	8	910
	1e. Planning coherent instruction	0	6	87	7	923
	1f. Designing ongoing formative assessments	0	22	74	3	901
2: Classroom environment	2a. Creating a learning environment of respect and rapport	0	9	73	17	927
	2b. Establishing a culture for learning*	0	12	74	13	1,069
	2c. Managing classroom procedures	0	9	80	11	929
	2d. Managing student behavior*	1	13	74	12	1,068
	2e. Organizing physical space	0	5	86	9	904
3: Teaching and learning	3a. Communicating with students	0	5	84	10	920
	3b. Using questioning and discussion techniques*	0	37	57	5	1,069
	3c. Engaging students in learning*	0	17	73	9	1,069
	3d. Using assessment to inform instruction*	0	25	68	6	1,069
	3e. Demonstrating flexibility and responsiveness	0	7	88	5	903
	3f. Assessment results and student learning*	1	21	74	4	1,067
	3g. Implementing lessons equitably*	0	9	84	6	1,068
4: Professional responsibilities	4a. Reflecting on teacher and student learning*	0	8	84	9	1,068
	4b. Systems for managing student data*	0	22	69	8	1,068
	4c. Communicating with families*	0	15	70	14	1,068
	4d. Participating in a professional community	0	6	84	10	907
	4e. Growing and developing professionally	0	7	84	9	900
	4f. Showing professionalism	0	3	84	13	904
Overall	Overall RISE rating	0	9	86	5	1,060

RISE is Research-based Inclusive System of Evaluation.

* One of the 12 power components.

a. 0 Unsatisfactory, 1 basic, 2 proficient, and 3 distinguished.

Source: Pittsburgh Public Schools data for the 2011/12 school year.

Table B4. Correlations between RISE components

		RISE component																								Overall RISE rating
Domain	RISE component	1a	1b	1c	1d	1e	1f	2a	2b	2c	2d	2e	3a	3b	3c	3d	3e	3f	3g	4a	4b	4c	4d	4e	4f	
1: Planning and preparation	1a. Demonstrating knowledge of content and pedagogy	1.00																								
	1b. Demonstrating knowledge of students*	.33	1.00																							
	1c. Setting instructional outcomes*	.41	.35	1.00																						
	1d. Demonstrating knowledge of resources	.40	.30	.32	1.00																					
	1e. Planning coherent instruction	.46	.33	.50	.32	1.00																				
	1f. Designing ongoing formative assessments	.31	.33	.35	.27	.35	1.00																			
2: Classroom environment	2a. Creating learning environment of respect & rapport	.30	.41	.33	.24	.32	.22	1.00																		
	2b. Establishing a culture for learning*	.35	.42	.44	.31	.40	.36	.48	1.00																	
	2c. Managing classroom procedures	.36	.35	.38	.32	.38	.32	.36	.40	1.00																
	2d. Managing student behavior*	.31	.37	.35	.26	.33	.26	.45	.48	.47	1.00															
	2e. Organizing physical space	.31	.22	.27	.27	.29	.19	.29	.25	.36	.26	1.00														
3: Teaching and learning	3a. Communicating with students	.35	.38	.39	.27	.41	.29	.39	.41	.34	.34	.24	1.00													
	3b. Using questioning and discussion techniques*	.35	.30	.36	.28	.32	.32	.29	.42	.31	.31	.16	.28	1.00												
	3c. Engaging students in learning*	.39	.39	.45	.29	.40	.31	.37	.57	.42	.45	.22	.35	.50	1.00											
	3d. Using assessment to inform instruction*	.34	.38	.40	.26	.35	.58	.29	.42	.35	.32	.20	.33	.39	.41	1.00										
	3e. Demonstrating flexibility and responsiveness	.32	.39	.30	.28	.35	.32	.35	.33	.30	.29	.29	.32	.26	.29	.30	1.00									
	3f. Assessment results and student learning*	.33	.39	.40	.25	.31	.49	.33	.39	.34	.34	.25	.29	.37	.40	.52	.35	1.00								
	3g. Implementing lessons equitably*	.29	.39	.42	.29	.39	.24	.37	.45	.37	.34	.25	.38	.32	.44	.31	.34	.33	1.00							

(continued)

Table B4. Correlations between RISE components *(continued)*

		RISE component																								Overall RISE rating
Domain	RISE component	1a	1b	1c	1d	1e	1f	2a	2b	2c	2d	2e	3a	3b	3c	3d	3e	3f	3g	4a	4b	4c	4d	4e	4f	
4: Professional responsibilities	4a. Reflecting on teacher and student learning*	.44	.34	.38	.35	.38	.33	.31	.37	.32	.30	.29	.36	.28	.36	.35	.33	.34	.30	1.00						
	4b. Systems for managing student data*	.30	.37	.40	.25	.29	.46	.24	.34	.30	.28	.23	.25	.27	.34	.52	.24	.46	.24	.35	1.00					
	4c. Communicating with families*	.16	.36	.26	.20	.19	.21	.28	.26	.25	.31	.22	.26	.20	.27	.24	.26	.29	.24	.28	.23	1.00				
	4d. Participating in a professional community	.30	.27	.28	.33	.28	.19	.24	.27	.26	.18	.22	.30	.22	.23	.21	.34	.24	.23	.37	.21	.28	1.00			
	4e. Growing and developing professionally	.27	.28	.25	.32	.26	.23	.25	.28	.30	.18	.23	.26	.21	.24	.26	.31	.23	.27	.38	.22	.28	.58	1.00		
	4f. Showing professionalism	.24	.23	.28	.24	.28	.21	.28	.27	.23	.17	.25	.31	.14	.18	.23	.23	.22	.27	.37	.22	.23	.43	.38	1.00	
Overall	Overall RISE rating	.45	.49	.53	.37	.47	.42	.48	.59	.44	.47	.30	.44	.41	.59	.51	.38	.55	.54	.50	.45	.34	.35	.36	.32	1.00

RISE is Research-based Inclusive System of Evaluation.

* One of the 12 power components.

Note: All correlations are statistically significant at the .05 level. Sample sizes range from 877 to 1,069 teachers.

Source: Pittsburgh Public Schools data for the 2011/12 school year.

Table B5. Distribution of 7Cs teacher ratings, by component

Component	Mean	Standard deviation	Minimum	Maximum
Cares	50.27	20.62	–8.64	99.00
Controls	50.10	20.69	–5.83	103.04
Clarifies	50.49	20.57	–8.48	98.24
Challenges	50.51	20.57	–4.07	96.97
Captivates	50.04	20.63	–1.16	97.29
Confers	50.61	20.48	–7.72	102.37
Consolidates	50.53	20.55	–7.09	100.39
7Cs composite[a]	50.36	17.79	–3.29	98.17

Note: The 7Cs ratings are from student surveys and are converted to approximately normal-curve-equivalent units. See appendix A for details. The sample size is 1,684 teachers.

a. Average across components.

Source: Pittsburgh Public Schools data for the 2011/12 school year.

Table B6. Correlations between RISE domains and overall ratings

	RISE domain				
RISE domain	1. Planning and preparation	2. Classroom environment	3. Teaching and learning	4. Professional responsibilities	Overall rating
1. Planning and preparation	1.00	.59	.82	.71	.80
2. Classroom environment	.63	1.00	.58	.50	.62
3. Teaching and learning	.75	.70	1.00	.82	.78
4. Professional responsibilities	.66	.54	.63	1.00	.65
Overall rating	.67	.65	.73	.60	1.00

RISE is Research-based Inclusive System of Evaluation.

Note: Domain variables are the averages of the components in each domain. Sample sizes range from 1,060 to 1,069 teachers. All correlations are statistically significant at the .05 level.

Source: Pittsburgh Public Schools data for the 2011/12 school year.

Table B7. Correlations between 7Cs components

	7Cs component							
7Cs component	Cares	Controls	Clarifies	Challenges	Captivates	Confers	Consolidates	Average
Cares	1.00	.59	.82	.71	.80	.76	.76	.90
Controls	.59	1.00	.58	.49	.62	.57	.50	.72
Clarifies	.82	.58	1.00	.82	.78	.79	.85	.93
Challenges	.71	.49	.82	1.00	.65	.69	.79	.85
Captivates	.80	.62	.78	.65	1.00	.71	.71	.87
Confers	.76	.57	.79	.69	.71	1.00	.75	.87
Consolidates	.76	.50	.85	.79	.71	.75	1.00	.89
Average 7Cs score	.90	.72	.93	.85	.87	.87	.89	1.00

Note: All correlations are statistically significant at the .05 level. Sample size is 1,684 teachers.

Source: Pittsburgh Public Schools data for the 2011/12 school year.

Table B8. Correlations between RISE composites and student characteristics

	Student characteristic					
RISE composite	Racial/ ethnic minority	Eligible for free or reduced-price lunch	English language learner	In special education	Gifted	Female
Average of all 24 components	**–.21**	**–.18**	.05	.02	**.14**	.01
Planning and preparation domain	**–.20**	**–.17**	.06	–.01	**.13**	.04
Classroom environment domain	**–.16**	**–.15**	**.06**	.02	**.13**	.02
Teaching and learning domain	**–.20**	**–.14**	.03	.03	**.11**	–.01
Professional responsibilities domain	**–.12**	**–.14**	.04	.02	**.10**	–.00
12 power components	**–.20**	**–.15**	.05	.05	**.12**	–.02
Danielson components	**–.20**	**–.18**	.05	.02	**.14**	.01

RISE is Research-based Inclusive System of Evaluation.

Note: Bold denotes statistically significant at the .05 level. For the professional responsibilities domain, sample size is 931 teachers; for all other measures, sample size is 932.

Source: Pittsburgh Public Schools data for the 2011/12 school year.

Table B9. Correlations between 7Cs components and student characteristics

	Student characteristic					
7Cs component	Racial/ ethnic minority	Eligible for free or reduced-price lunch	English language learner	Special education	Gifted	Female
Cares	**–.14**	**–.08**	.03	**.07**	.04	–.04
Controls	**–.25**	**–.26**	.03	**–.07**	**.21**	.05
Clarifies	**–.10**	–.04	.04	.03	**.06**	–.02
Challenges	**–.11**	–.05	.01	–.00	**.08**	.03
Captivates	**–.13**	**–.10**	**.07**	.01	**.09**	–.03
Confers	**–.10**	**–.09**	**.06**	.03	**.11**	–.04
Consolidates	–.04	.05	.05	**.13**	–.03	**–.11**
Average 7Cs score	**–.14**	**–.09**	.05	.03	**.09**	–.03

Note: Bold denotes statistically significant at the .05 level. Sample size is 1,535 teachers.

Source: Pittsburgh Public Schools data for the 2011/12 school year.

Table B10. Coefficient estimates for RISE composites regressed on all student characteristics

	Student characteristic					
RISE composite	Racial/ ethnic minority	Eligible for free or reduced-price lunch	English language learner	Special education	Gifted	Female
Average of all 24 components	**–0.20**	–0.28	–0.00	0.00	0.10	–0.17
Planning and preparation domain	**–0.27**	–0.25	0.06	–0.04	0.03	–0.02
Classroom environment domain	–0.12	–0.15	–0.04	0.06	**0.49**	–0.23
Teaching and learning domain	**–0.30**	–0.24	–0.08	0.02	–0.01	–0.16
Professional responsibilities domain	–0.09	**–0.47**	0.05	–0.02	–0.05	–0.28
12 power components	**–0.27**	–0.24	0.10	0.05	0.15	–0.27
Danielson components	**–0.20**	–0.29	–0.01	0.01	0.12	–0.19

RISE is Research-based Inclusive System of Evaluation.

Note: Sample size is 1,014 teachers, except for the professional responsibilities domain, for which it is 1,013. Bold denotes statistically significant at the 0.05 level.

Source: Pittsburgh Public Schools data for the 2011/12 school year.

Table B11. Coefficient estimates for 7Cs components regressed on all student characteristics

	Student characteristic					
7Cs component	Racial/ ethnic minority	Eligible for free or reduced-price lunch	English language learner	Special education	Gifted	Female
Cares	–3.69	**–18.26**	15.47	**15.93**	–6.58	–16.87
Controls	**–9.72**	**–19.93**	–4.06	4.92	8.11	–3.64
Clarifies	–0.02	–8.72	14.45	8.23	1.28	**–20.66**
Challenges	–5.22	2.59	–6.01	7.69	17.25	–1.53
Captivates	2.31	**–20.04**	32.87	3.71	–7.13	**–27.78**
Confers	–2.88	–13.53	6.85	10.13	12.86	–7.91
Consolidates	–1.48	–1.45	21.01	**11.38**	1.86	–17.00
Average 7Cs score	–2.96	–11.33	11.51	8.85	3.95	–13.63

Note: Student characteristics describe students of a given teacher during the 2011/12 school year. Bold denotes statistically significant at the 0.05 level. Sample size is 1,536 teachers.

Source: Pittsburgh Public Schools data for the 2011/12 school year.

Table B12. Correlations between value-added measures and student characteristics

	Student characteristic						
Value-added measure	Racial/ ethnic minority	Eligible for free or reduced-price lunch	English language learner	Special education	Gifted	Female	Sample size
English language arts	–.02	**–.16**	–.03	**–.17**	**.24**	**.14**	310
Math	.05	.07	.07	.04	–.10	–.03	250
Science	**–.25**	**–.28**	.07	–.04	**.26**	.00	131
Social studies	.12	.03	.02	.16	–.01	–.05	69
All grade levels	–.02	–.06	.04	–.06	.07	.06	641
Elementary school	–.03	–.11	–.06	–.05	**.20**	.00	181
Middle school	–.02	–.04	.07	–.10	.05	.05	303
High school	–.08	–.07	.09	.01	.09	.03	213
Pennsylvania System of School Assessment	.00	–.06	–.01	–.09	**.10**	.06	447
Curriculum-based assessment	**-.11**	**–.12**	.07	–.07	**.10**	.04	454

Note: Bold denotes statistically significant at the .05 level.

Source: Pittsburgh Public Schools data for the 2011/12 school year. Value-added data are from Johnson et al. (2012) and are for 2008–11.

Table B13. Estimated effects of teacher experience on value-added measure estimates

	Years of teacher experience					
Value-added measure	2–3	4–8	9–20	More than 20	R-squared	Sample size
English language arts	**16.75**	**20.00**	**22.32**	**20.75**	.073	169
Math	10.33	10.33	**19.91**	**17.37**	.111	137
Science	5.04	5.12	8.14	20.84	.086	75
Social studies	–4.42	1.39	6.10	24.39	.104	45
Elementary school	12.63	**14.12**	**21.66**	**31.48**	.216	101
Middle school	14.60	18.56	**23.91**	21.45	.065	172
High school	4.41	7.58	11.30	**15.91**	.057	120
Pennsylvania System of School Assessment	**15.01**	**19.02**	**21.99**	**22.86**	.070	244
Curriculum-based assessment	6.25	6.98	**12.59**	**13.03**	.040	249
All grade levels	**10.02**	**13.65**	**17.97**	**20.65**	.066	353

Note: The comparison group is teachers with less than two years of experience. Bold denotes statistically significant at the .05 level.

Source: Pittsburgh Public Schools data for the 2011/12 school year.

Table B14. RISE performance by availability of 7Cs and value-added measure data

	Overall RISE rating		Average of 12 power components		Average of all 24 components		
Data availability	Group mean	Number of teachers	Group mean	Number of teachers	Group mean	Number of teachers	Total number of teachers
No 7Cs or Value-added data	1.98	175	1.95	165	1.97	117	365
7Cs data only	**1.93**	513	1.91	506	1.97	371	982
Value-added data only	1.92	12	1.74	10	**1.74**	7	33
Both	1.98	391	1.93	386	1.99	312	702
Total	1.96	1,091	1.92	1,067	1.98	807	2,082

RISE is Research-based Inclusive System of Evaluation.

Note: Overall RISE rating is the rating on the overall RISE measure, not an average of the components. Bold denotes that the mean for this group differs statistically from the group with both 7Cs and Value-added data at the .05 level. Results for the average of the 12 power components and of all 24 components are limited to the teachers with all 12 power components and all 24 RISE components.

Source: RISE and 7Cs data are from Pittsburgh Public Schools for the 2011/12 school year; Value-added data are from Johnson et al. (2012) and are for 2009–12.

Appendix C. Principal components analyses

To assist Pittsburgh Public Schools' efforts to improve the Research-based Inclusive System of Evaluation (RISE) metric, this report develops several composite measures of RISE. Measures based on averages were described in the main body of the report. This appendix describes summaries calculated using principal components analysis, which was used by Kane and Staiger (2012) in their analyses of measures of teachers' professional practice. The measure obtained using principal components analysis proved very similar to an unweighted average of the components because the weights used were very similar across components.

The technical advisory board for this project recommended consideration of factor analysis as an alternative to principal components analysis. Factor analysis is designed to identify causal mechanisms behind observed measures. Analyses using factor analyses produced no evidence that the resulting variables were better (or worse) for summarizing the data. Since the main goal was to identify a small number of composite measures that explain the variance in RISE ratings rather than to conduct a theory-based analysis of the underlying structure of the variables, this report used principal components analysis.

Principal components analysis was used to identify a small set of variables that summarize the variation found in the larger set of all RISE components. The analysis identified four variables that capture the variation found in larger sets of RISE components (all 24 and the 12 power components). Variables identified using principal components analysis are commonly called principal components. For this study they are the weighted sums of the RISE components used.

Composite measures based on the 12 power components perform about as well as those based on all 24 components in internal consistency and in their relationships to value-added measures (VAMs). Therefore, this appendix focuses on four composite measures of the RISE power components. Each composite measure, or principal component, is a weighted sum of the RISE power components. The weights used in creating the four principal components are shown in table C1. The power components with the largest weights suggest that the four principal components might be characterized as follows:

- *Principal component 1.* Teacher's overall effectiveness (approximately equal weights for all components).
- *Principal component 2.* Teacher's use of data and assessments (largest weights for systems for managing student data, use of assessment to inform instruction, assessment results, and student learning).
- *Principal component 3.* Teacher's knowledge of students and their families (largest weights for communicating with families and demonstrating knowledge of students).
- *Principal component 4.* Teacher's focus on student learning (largest weight for using questioning and discussion techniques).

The four variables identified through principal components analysis and based on weighted averages of the 12 RISE power components (see table C1) explain more than half the variation in the power components. In addition, about half (52 percent) of teachers have performance ratings that can be distinguished from the mean value of the first principal component based on the 12 RISE power components. The weights used for this calculation

Table C1. Weights for the 12 RISE power components for principal components analysis

Power component	Principal component 1	Principal component 2	Principal component 3	Principal component 4
1b. Demonstrating knowledge of students	.29	.00	.33	.04
1c. Setting instructional outcomes	.30	.01	–.10	–.39
2b. Establishing a culture for learning	.33	–.28	–.13	.01
2d. Managing student behavior	.28	–.29	.16	.27
3b. Using questioning and discussion techniques	.27	–.18	–.38	.38
3c. Engaging students in learning	.33	–.29	–.21	.10
3d. Using assessment to inform instruction	.31	.41	–.21	.15
3f. Assessment results and student learning	.30	.34	–.06	.19
3g. Implementing lessons equitably	.28	–.35	–.02	–.40
4a. Reflecting on teacher and student learning	.27	.12	.18	–.57
4b. Systems for managing student data	.28	.55	–.06	–.03
4c. Communicating with families	.22	–.01	.75	.28

RISE is Research-based Inclusive System of Evaluation.

Note: Weights were calculated for the sample of 1,067 teachers with complete data for all 12 RISE power components.

Source: Pittsburgh Public Schools data for the 2011/12 school year.

were estimated using a randomly selected half of the data. These weights were then used to calculate the percentage of teachers with statistically significant results using the other half. This ensured that the results were not biased upward by the fact that principal components analysis sets weights based on how strongly the components are correlated with each other.

The correlations with VAM estimates are very similar for the principal components and for an unweighted average of the RISE components. More precisely, the correlation of the overall VAM estimates is about .23 for the first principal component (the one with weights that varied the least across components) and about .24 for an unweighted average of the 12 power components. The results are almost identical for the three-year average of the VAM estimates ending in 2012 rather than 2011. Also, the *p*-value is about .02 for a regression of the VAM estimates ending in 2011 on the first principal component and about the same for an unweighted average of the 12 power components; both remain statistically significant after controlling for grade and subject dummy variables.

These results indicate that, for the RISE measure, principal components analysis provides a reasonably concise and reliable way to distinguish among teacher performance levels, though it would be preferable to have data on multiple years of RISE data to obtain a better measure of reliability.

Appendix D. Correlating the Research-based Inclusive System of Evaluation and 7Cs ratings with partly concurrent value-added measure estimates

Because true teacher effectiveness likely changes over time, examining correlations of measures from different school years could underestimate the true correlations of the Research-based Inclusive System of Evaluation (RISE) and 7Cs survey ratings with value-added measure (VAM) estimates. To address this concern, correlations were calculated for the 2011/12 RISE and 7Cs ratings with partly concurrent VAM estimates for 2009–12 (rather than for 2008–11).[9]

Surprisingly, relationships between RISE ratings and VAM estimates for partly concurrent years (table D1) are weaker than the relationships between RISE ratings and VAM estimates for nonoverlapping years (see table 6 in main report). Using VAM estimates that include the year of the RISE rating produces smaller correlations and fewer that are statistically significant. This comparison is misleading, however, as the two sets of analyses involve different samples of teachers. More specifically, many teachers in the VAM correlation analyses for the partly concurrent years are first-year teachers (who had no VAM estimates for 2008–11 and therefore could not be included in the original correlational analysis).[10] The correlations of their RISE ratings with VAM estimates might therefore be lower than for other teachers. When the sample is limited to teachers with VAM estimates for both sets of years, the correlations are far more similar.

For 7Cs ratings, in contrast, the relationship is much stronger with partly concurrent VAM estimates than with the nonoverlapping year estimates. For example, 83 percent of the partly concurrent VAM estimates are statistically significant (table D2), but only 56 of the nonoverlapping years correlations are statistically significant (see table 7 in main report).[11] In addition, the correlation between the composite 7Cs rating and the composite VAM estimate is only .15 for the nonoverlapping years compared with .22 for the partly concurrent VAM correlations with 7Cs. This differential holds when the sample is limited to teachers with VAM estimates in both sets of years, suggesting that the relationship is stronger when based on overlapping years for the two measures.

At first glance, it seems surprising that the relationships to partly concurrent and to nonoverlapping VAM estimates differ for RISE (higher or equal correlation with nonoverlapping VAM estimates) and 7Cs (higher correlation with partially concurrent VAM estimates). There are several possible explanations. Principals may focus more on true long-run effectiveness when making judgments for the RISE measures and may even use information collected in previous years, perhaps explaining the relatively consistent correlations of RISE ratings with VAM estimates for nonoverlapping years for the same set of teachers. The lower correlation when first-year teachers are included may reflect principals' uncertainty about how to rate such teachers. In contrast, students filling out the 7Cs surveys are likely to be highly focused on true current-year effectiveness.

Table D1. Correlations between RISE ratings for 2011/12 and partly concurrent value-added measure estimates for 2009–12

		Value-added measure									
RISE domain	RISE component	English language arts	Math	Science	Social studies	All grade levels	Elementary school	Middle school	High school	Pennsylvania System of School Assessment	Curriculum-based assessment
1: Planning and preparation	1a. Demonstrating knowledge of content and pedagogy	.04	–.17	.04	–.12	–.01	–.05	.06	–.17	.03	–.08
	1b. Demonstrating knowledge of students*	.13	.12	.07	.26	**.10**	.09	.08	**.21**	.09	**.13**
	1c. Setting instructional outcomes*	.03	–.06	.14	.08	.04	–.03	.05	.06	.08	.02
	1d. Demonstrating knowledge of resources	.11	–.07	.13	–.04	.07	.09	.10	–.06	.10	.00
	1e. Planning coherent instruction	.12	–.02	.12	–.17	.06	.06	.12	–.02	.12	.03
	1f. Designing ongoing formative assessments	.04	.04	.17	.17	.09	–.01	.15	.10	.08	.08
2: Classroom environment	2a. Creating learning environment of respect and rapport	**.18**	.13	.10	.10	**.12**	.17	.13	–.02	**.16**	.09
	2b. Establishing a culture for learning*	**.18**	.08	.04	.22	**.12**	**.17**	.15	.03	**.20**	.03
	2c. Managing classroom procedures	**.18**	.05	.01	**.27**	**.12**	**.18**	.05	.09	**.13**	.04
	2d. Managing student behavior*	**.21**	.07	.10	.16	**.14**	.15	**.16**	.09	**.18**	.10
	2e. Organizing physical space	.04	**–.26**	.05	–.11	–.03	.04	–.08	–.16	.03	–.12
3: Teaching and learning	3a. Communicating with students	.09	.09	.21	.15	.10	.08	.14	.00	**.15**	.09
	3b. Using questioning and discussion techniques*	.11	.09	.19	**.29**	**.14**	.14	**.18**	.05	**.17**	**.12**
	3c. Engaging students in learning*	.07	.08	.08	.16	**.13**	.09	.12	.11	**.16**	.06
	3d. Using assessment to inform instruction*	.06	.06	.01	.17	.08	–.05	**.18**	.11	.08	.10
	3e. Demonstrating flexibility and responsiveness	.11	.05	.17	.03	.08	.17	.05	.02	.09	.08
	3f. Assessment results and student learning*	.00	.05	.15	.24	**.12**	.06	.11	.11	.11	.08
	3g. Implementing lessons equitably*	.13	.14	.14	.01	**.13**	**.18**	.08	.09	**.19**	–.00
4: Professional responsibilities	4a. Reflecting on teacher and student learning*	.12	.04	–.05	.21	.06	.08	.09	.01	.11	.03
	4b. Systems for managing student data*	.02	.02	–.01	.16	.03	–.04	.09	.07	.01	.07
	4c. Communicating with families*	.09	–.04	.19	.24	.08	.06	.12	.11	.07	.11
	4d. Participating in a professional community	.09	.02	.11	.04	.03	.08	.01	.01	.05	.04
	4e. Growing and developing professionally	**.16**	–.00	–.02	–.13	.00	.12	–.09	.02	.07	–.04
	4f. Showing professionalism	.02	–.05	.14	–.23	.02	.06	–.03	–.03	.07	–.03
Averages	All 24 components	**.15**	.05	.16	.25	**.14**	.12	**.16**	.12	**.18**	.09
	12 power components	**.16**	.08	.15	**.30**	**.16**	.12	**.21**	.14	**.20**	.12
	Danielson components	**.15**	.04	.15	.25	**.14**	.11	**.16**	.11	**.18**	.10
Overall		.12	.11	**.22**	.03	**.15**	**.19**	.04	.13	**.19**	.03

RISE is Research-based Inclusive System of Evaluation.

* One of the 12 power components.

Note: Sample sizes are shown in table D3. Bold denotes statistically significant at the .05 level.

Source: RISE data are from Pittsburgh Public Schools for the 2011/12 school year; Value-added data are from Johnson et al. (2012) and are for 2009–12.

Table D2. Correlations between 7Cs ratings and partly concurrent value-added measure estimates

	Value-added measure									
7Cs domain	English language arts	Math	Science	Social studies	All grade levels	Elementary school	Middle school	High school	Pennsylvania System of School Assessment	Curriculum-based assessment
Cares	.11	**.13**	.13	**.27**	**.14**	.02	**.22**	**.22**	**.12**	**.20**
Controls	**.21**	**.22**	**.23**	**.36**	**.24**	**.20**	**.23**	**.30**	**.19**	**.26**
Clarifies	**.19**	**.18**	.15	**.32**	**.21**	.10	**.26**	**.26**	**.19**	**.22**
Challenges	**.30**	**.22**	.14	**.36**	**.26**	**.21**	**.25**	**.31**	**.26**	**.21**
Captivates	**.12**	.08	**.28**	**.28**	**.16**	.03	**.22**	**.25**	**.13**	**.22**
Confers	**.16**	.07	.12	**.44**	**.17**	.05	**.21**	**.28**	**.15**	**.21**
Consolidates	**.19**	**.19**	.15	**.29**	**.21**	.10	**.25**	**.29**	**.19**	**.21**
Average 7Cs rating	**.21**	**.18**	**.20**	**.37**	**.22**	.12	**.26**	**.30**	**.20**	**.24**

Note: 7Cs measures are based on student survey data. Sample sizes are shown in table D4. Bold denotes statistically significant at the .05 level.

Source: 7Cs data are from Pittsburgh Public Schools for the 2011/12 school year; Value-added data are from Johnson et al. (2012) and are for 2009–12.

Table D3. Sample sizes used to estimate correlations between RISE and current year value-added measures

		Value-added measure									
RISE domain	RISE component	English language arts	Math	Science	Social studies	All grade levels	Elementary school	Middle school	High school	Pennsylvania System of School Assessment	Curriculum-based assessment
1: Planning and preparation	1a. Demonstrating knowledge of content and pedagogy	169	130	67	49	357	121	153	119	246	243
	1b. Demonstrating knowledge of students*	189	146	81	55	397	136	165	137	269	271
	1c. Setting instructional outcomes*	189	146	81	55	397	136	165	137	269	271
	1d. Demonstrating knowledge of resources	168	130	66	49	355	120	151	120	244	242
	1e. Planning coherent instruction	169	132	66	49	356	122	156	117	247	244
	1f. Designing ongoing formative assessments	167	129	66	49	353	120	151	117	243	240
2: Classroom environment	2a. Creating learning environment of respect and rapport	172	132	67	50	363	123	159	119	251	248
	2b. Establishing a culture for learning*	189	146	81	55	397	136	165	137	269	271
	2c. Managing classroom procedures	170	128	68	54	360	122	154	121	246	245
	2d. Managing student behavior*	189	146	81	55	397	136	165	137	269	271
	2e. Organizing physical space	166	128	64	50	349	119	150	116	242	237
3: Teaching and learning	3a. Communicating with students	170	131	67	50	360	122	155	120	247	245
	3b. Using questioning and discussion techniques*	189	146	81	55	397	136	165	137	269	271
	3c. Engaging students in learning*	189	146	81	55	397	136	165	137	269	271
	3d. Using assessment to inform instruction*	189	146	81	55	397	136	165	137	269	271
	3e. Demonstrating flexibility and responsiveness	163	130	64	49	348	116	152	116	240	239
	3f. Assessment results and student learning*	188	146	81	55	396	135	165	137	268	271
	3g. Implementing lessons equitably*	188	146	81	55	396	135	165	137	268	271
4: Professional responsibilities	4a. Reflecting on teacher and student learning*	188	146	81	55	396	135	165	137	268	271
	4b. Systems for managing student data*	188	146	81	55	396	135	165	137	268	271
	4c. Communicating with families*	188	146	81	55	396	135	165	137	268	271
	4d. Participating in a professional community	170	129	68	51	358	120	152	122	245	245
	4e. Growing and developing professionally	166	129	65	49	351	120	151	116	243	238
	4f. Showing professionalism	167	129	66	48	352	120	151	117	244	239
Averages	All 24 components	189	146	81	55	397	136	165	137	269	271
	12 power components	189	146	81	55	397	136	165	137	269	271
	Danielson components	189	146	81	55	397	136	165	137	269	271
Overall	Overall RISE rating	186	145	81	55	393	134	163	137	265	269

RISE is Research-based Inclusive System of Evaluation.

* One of the 12 power components.

Source: RISE data are from Pittsburgh Public Schools for the 2011/12 school year; Value-added data are from Johnson et al. (2012) and are for 2009–12.

Table D4. Sample sizes used to estimate correlations between 7Cs and current year value-added measures

	Value-added measure									
7Cs component	English language arts	Math	Science	Social studies	All grade levels	Elementary school	Middle school	High school	Pennsylvania System of School Assessment	Curriculum-based assessment
All	341	278	147	69	702	258	299	198	513	447

Note: Sample sizes are the same for all 7Cs components, so only one row is needed.

Source: 7Cs data are from Pittsburgh Public Schools for the 2011/12 school year; Value-added data are from Johnson et al. (2012) and are for 2009–12.

Appendix E. Checking Research-based Inclusive System of Evaluation ratings within and between schools

It is important to know whether principals rate teachers within schools in ways that align with teachers' true effectiveness, in part by assessing their relationship with ratings based on the value-added measure (VAM) and the 7Cs surveys. This appendix presents results from regressions of the average of all 24 components of the Research-based Inclusive System of Evaluation (RISE) on the average VAM estimate, the 7Cs average, and school dummy variables. Thus, the coefficient estimates for the VAM and 7Cs variables are driven by within-school variation (combined across all schools). Both the VAM and 7Cs composite measures have positive and statistically significant coefficients, suggesting that, on average, principals' rankings of teachers within schools are generally consistent with rankings based on VAM and 7Cs ratings.

It is also important to examine whether principals' RISE ratings of teachers are consistent across schools. This can be done by replacing the school dummy variables in the regression with school averages of the VAM and 7Cs average variables to see whether RISE ratings are more consistent with VAM estimates and 7Cs ratings within or between schools. The coefficients on the teacher-level VAM and 7Cs variables are still identified by the within-school variation in RISE ratings. The coefficient estimates on the school average variables show whether an increase in those variables at the school level has any additional impact on RISE ratings beyond the impacts indicated by the coefficients on the teacher-level variables. Thus, positive and statistically significant coefficients on the school average variables would suggest that principals rate teachers in ways that are more highly correlated with VAM differences across schools than within schools. This method is similar to the one used by Rockoff and Speroni (2011) to analyze the consistency of teacher evaluations across raters in New York City schools.[12]

The coefficients on the teacher-level VAM and 7Cs variables are positive and almost all are statistically significant (table E1). This suggests that principals are rating their teachers in ways that align with VAM estimates and 7Cs ratings within their schools. The coefficients on the school-average VAM variable are also positive and statistically significant for two of the RISE variable regressions but not the coefficients on the school average 7Cs variable. That suggests that principals are rating teachers in ways that are more clearly aligned with the between-school variation in VAM than the within-school variation. There are a number of possible explanations for this result.[13] One is that principals might view teachers as teams and believe that they are jointly responsible for growth observed at the school level. Another explanation relates to principals' prior knowledge of schoolwide VAM, which could have influenced their RISE ratings across the school. In contrast, the principals did not have access to teacher-level VAM data, which might explain why their ratings are more strongly related to the school-level VAM than to individual teacher's VAM within schools.

As noted above, the coefficients on the school-average 7Cs variables are not statistically significant in any of the regressions in table E1, each of which controls for the teacher averages of the 7Cs variables. Statistically significant coefficients might be found if students rated teachers differently within schools than across schools relative to how principals rate teachers. The analysis thus found no evidence of such differential rating of teachers by students compared to by principals in these regressions. This result is also consistent

Table E1. Regression of RISE composite ratings on teacher and school value-added average and 7Cs average

	Teacher averages		School averages	
RISE composite	Value-added measure estimates	7Cs	Value-added measure estimates	7Cs
Overall RISE rating	**0.0021**	**0.0039**	0.0038	–0.0034
Average of all 24 components	**0.0021**	**0.0037**	**0.0040**	0.0015
Domain				
1. Planning and preparation	0.0013	**0.0030**	0.0039	0.0025
2. Classroom environment	**0.0030**	**0.0063**	**0.0057**	–0.0044
3. Teaching and learning	**0.0033**	**0.0036**	0.0036	0.0029
4. Professional responsibilities	0.0006	**0.0023**	0.0031	0.0047
12 power components	**0.0025**	**0.0045**	0.0043	0.0020

RISE is Research-based Inclusive System of Evaluation.

Note: Bold indicates statistically significant at .the 05 level. The 7Cs and value-added measure variables are in normal-curve-equivalent units. Sample size ranges from 326 to 329 teachers.

Source: RISE and 7Cs data are from Pittsburgh Public Schools for the 2011/12 school year; Value-added data are from Johnson et al. (2012) and are for 2008–11.

with the fact that principals did not have access to the 2011/12 7Cs data at the school or teacher levels when they assigned their 2011/12 RISE ratings. (7Cs was not administered districtwide before the 2011/12 school year.)

As a final step, averages of the residuals from the regressions in table E1 by school were used to check rater consistency (the differences between the observed RISE ratings and those predicted based on the VAM and 7Cs results). For each regression shown in table E1 the residuals differed across schools in statistically significant ways. This suggests that while RISE ratings do align somewhat with the VAM estimates and the 7Cs ratings, RISE ratings also vary across schools in ways that are not explained by VAM estimates and 7Cs ratings. This may mean that RISE is capturing teacher skills that are not captured by the other two measures. Alternatively, the additional variation in ratings by school might mean that principals are not using the same teacher rating standards across schools.

Notes

1. Correlations of each RISE component with each of the other components range from .14 to .58, with an average of .32 (table B4 in appendix B).
2. Correlations of the RISE domain composite ratings (average component rating within each domain) range from .50 to .82, with an average of .67 (table B6 in appendix B). Correlations of the 7Cs components with each other are similar to those for the RISE domains, ranging from .49 to .85, with an average of .70 (table B7 in appendix B).
3. Results describing the relationships of multiple RISE and 7Cs measures with student characteristics are in tables B8 and B9 in appendix B. Regressions of the RISE and 7Cs measures using all student characteristics in each regression suggested similar patterns; tables B10 and B11.
4. The regressions reported in tables 4 and B13 differ because of missing VAM and RISE values and because table B13 omits teacher characteristics other than experience. The results are qualitatively similar when limited to the sample of teachers with both VAM and RISE data and using only the experience categories as regressors.
5. The district was considering using component 3f as a substitute for VAM estimates for teachers lacking VAM data.
6. Regression results suggested that controls for student characteristics have no impact on the relationships between RISE ratings and VAM estimates.
7. In the future, the district aims to give "instructional teacher leaders" in each school a larger role in the RISE process. By 2013/14 teachers may receive separate RISE ratings from principals and instructional teacher leaders.
8. The 7Cs survey data are scaled using modified normal-curve-equivalent units to help guard against the influence of outliers (extremely large or small value data points) and deviations from symmetry in the original data. Normal curves are bell-shaped and perfectly symmetrical. A standard normal-curve-equivalent variable is a linear function of a *z*-score (*z*-score*21.06+50), while a *z*-score is a standardized linear transformation of the underlying data. One problem with the standard normal-curve-equivalent variable is that outliers may have undue influence on results based on those data. Another problem is that the distribution of values in the original data may not be perfectly bell-shaped, as the normal-curve-equivalent formula assumes. Both of these problems could matter when the 7Cs results are combined with other measures of teacher performance to calculate an overall evaluation score. The modified normal-curve-equivalent variable used for the 7Cs data mitigates these problems by being centered around the median instead of the mean and by using top and bottom coding at 1 and 99. For this study the data were re-normed by grade, so the resulting values in the re-normed data occasionally extend outside of the 1–99 range.
9. The correlation between the VAM estimates covering 2008–11 and the later year measure (covering 2009–12) is around .61. This is much higher than correlations between value-added results based on single-year estimates because these estimates overlap for two of the three years covered.
10. Also, the 2008–11 data have no grade 3 teachers while the 2009–12 data do.
11. When the 7Cs sample was limited to teachers with at least 12 RISE ratings, the sample size was about half as large, and there were far fewer statistically significant correlations.
12. See, for example, table 3 in their paper. Their methods also differ from the methods of this report in several potentially important ways. For example, they used student achievement as the outcome and controlled for various student and teacher

characteristics on the right side of the equation. The key right-side variables describe teacher observation ratings at the teacher and rater levels. In contrast, this report uses teacher observation ratings as the outcome and student achievement/VAM estimates on the right side.

13. The result presented here differs from that found by Rockoff and Speroni (2011), who found a stronger relationship between value-added and observation ratings for individual raters than between raters.

References

Bill & Melinda Gates Foundation. (2010). *Learning about teaching: Initial findings from the Measures of Effective Teaching project* (Policy Brief, MET Project). Seattle, WA: Author. http://eric.ed.gov/?id=ED528388

Bill & Melinda Gates Foundation. (2012). *Asking students about teaching: Student perception surveys and their implementation* (Policy and Practice Brief, MET Project). Seattle, WA: Author.

Danielson, C. (2013). *The Framework for Teaching Evaluation Instrument.* Princeton, NJ: The Danielson Group.

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

Johnson, M., Lipscomb, S., Gill, B., Booker, K., & Bruch, J. (2012). *Value-added models for the Pittsburgh Public Schools.* Cambridge, MA: Mathematica Policy Research. http://eric.ed.gov/?id=ED531790

Kane, T. J. (2012). Capturing the dimensions of effective teaching. *Education Next, 12*(4), 35–41. http://eric.ed.gov/?id=EJ994599

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains.* Seattle, WA: Bill & Melinda Gates Foundation. http://eric.ed.gov/?id=ED540961

Kimball, S., White, B., Milanowski, A., & Borman, G. (2004). Examining the relationship between teacher evaluation and student assessment results in Washoe County. *Peabody Journal of Education, 79*(4), 54–78. http://eric.ed.gov/?id=EJ683109

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

Milanowski, A. T. (2011). *Validity research on teacher evaluation systems based on the Framework for Teaching.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA. http://eric.ed.gov/?id=ED520519

Rockoff, J., & Speroni, C. (2011). Subjective and objective evaluations of teacher effectiveness: Evidence from New York City. *Labour Economics, 18*(5), 687–696.

Wasserman, J. D., & Bracken, B. A. (2003). Psychometric characteristics of assessment procedures. In I. B. Wiener, J. R. Graham, & J. A. Naglieri (Eds.), *Handbook of Psychology, Volume 10: Assessment Psychology* (pp. 43–66). Hoboken, NJ: John Wiley and Sons, Inc.