MATHEMATICA Policy Research

REPORT

Value-Added Models for the Pittsburgh Public Schools, 2012-13 School Year

May 5, 2014

Dana Rotz
Matthew Johnson
Brian Gill

Submitted to:
Pittsburgh Public Schools
Office of Research, Assessment, and Accountability
341 S. Bellefield Ave.
Pittsburgh, PA 15214
Project Officer: Mary Wolfson
Contract Number: 0E9347

Submitted by:
Mathematica Policy Research
955 Massachusetts Avenue
Suite 801
Cambridge, MA 02139
Telephone: (617) 491-7900
Facsimile: (617) 491-8044
Project Director: Brian Gill
Reference Number: 06723.300

This page has been left blank for double-sided copying.

ACKNOWLEDGMENTS

The authors would like to thank the many staff—too many to name here—of Pittsburgh Public Schools and the Pittsburgh Federation of Teachers who provided input and assistance over the last four years that this work has been underway. Although we cannot name everyone, we especially want to thank Mary Wolfson, Josh Aderholt, Veronica Amundson, Sam Franklin, Eddy Jones, and Bill Hileman for their support for the work. Our technical advisory group, consisting of Howard Nelson, John Tyler, Drew Gitomer, and Cory Koedel, also made helpful suggestions along the way.

Several staff at Mathematica Policy Research in addition to the authors contributed to this report. We thank Eric Isenberg, Duncan Chaplin, Hanley Chiang, and John Deke for their technical suggestions, and Juha Sohlberg for organizing the data and estimating the value-added models. Brynn Hagen ably led the process for producing individual value-added reports for teachers and schools, assisted by Mickey McCauley and Julia Cohen. Margaret Hallisey and Autumn Parker formatted the report to prepare it for public release.

This page has been left blank for double-sided copying.

GLOSSARY

Confidence interval	A confidence interval is the range of values (for example, around a teacher VAM estimate) within which the true value is expected to lie.
Correlation coefficient	The correlation coefficient measures the extent to which two variables are linearly related. Correlations near one indicate that values of the second variable are likely to increase when values of the first variable increase. Correlations close to zero indicate that the two variables are largely independent of each other.
Dosage	Dosage is the fraction of a student's instruction in a particular subject and academic year for which a specific school or teacher is responsible.
Mean standard error	The mean standard error is the average error around a set of estimates, such as around all teacher VAM estimates. Smaller standard errors translate into more precise estimates.
R-squared	The r-squared of a model is a measure of its goodness of fit to the data. High values of r-squared suggest that the model is likely to predict future outcomes well.
Sampling error	Sampling error is the error from chance differences in the characteristics of the sample studied relative to the overall population.
Shrinkage	Shrinkage is a post-estimation process that helps to ensure that teachers or schools with imprecise estimates are not over-represented among high-performers and low-performers.
Standard deviation	A standard deviation (SD) measures how much variability from average is in the data. According to a bell curve, the 84th percentile is one SD above average. The 98th percentile is two SDs above average.
Statistically significant	An estimate is statistically significant if the values in its corresponding confidence interval are either all above or all below zero. Larger confidence intervals (such as a 95 percent interval rather than a 90 percent interval) increase the chance that values overlap with zero, but strengthen the inference when values do not overlap with zero.
Value-added model	A value-added model is a statistical framework for identifying the individual contributions of teachers or schools to the achievement of their students.

This page has been left blank for double-sided copying.

CONTENTS

TABLES

This page has been left blank for double-sided copying.

FIGURES

This page has been left blank for double-sided copying.

I. INTRODUCTION

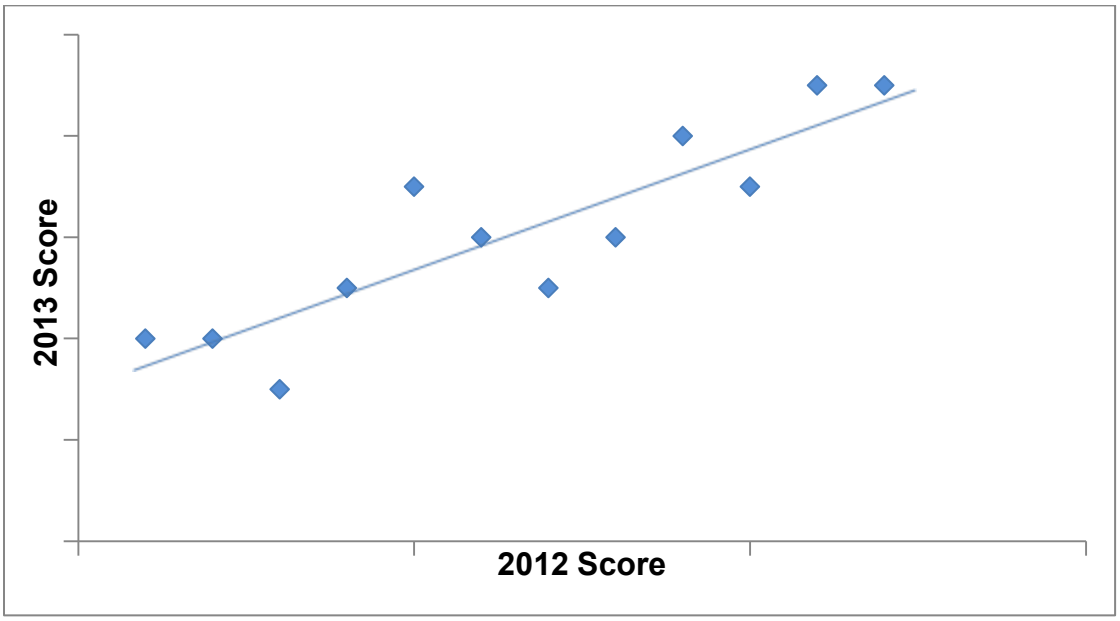
At the request of Pittsburgh Public Schools (PPS) and the Pittsburgh Federation of Teachers (PFT), Mathematica has developed value-added models that aim to estimate the contributions of individual teachers and schools to the achievement of their students. Our work in estimating value-added measures (VAMs) in Pittsburgh supports the larger, joint efforts of PPS and the PFT to empower effective teachers through evaluation, professional development, and compensation. Pittsburgh's value-added models use not only state assessments but also course-specific assessments, student attendance, and course completion rates, thereby aiming to produce estimates of the contributions of teachers and schools that are fair, valid, reliable, and robust. This report summarizes Mathematica's current use of value-added models to assess educational quality, updating our 2012 report on the same topic (Johnson et al. 2012).

A VAM provides a better indication of effectiveness than average score levels or the rate of student proficiency because it accounts for students' prior achievement and other factors that are outside the control of teachers or schools (Meyer 1997). The process of estimating a teacher or school value-added model can be conceptualized as occurring in two steps. In the first step, the model makes a prediction about an outcome of interest, typically a student's assessment score in a subject, based on factors that include students' prior achievement and other characteristics of students and their peers. The influence of these factors in the prediction is determined empirically, by examining relationships between the factors and student achievement in a large population of students. Each student's own prior achievement is generally the most important element in the prediction. These predictions represent what the students would be expected to achieve if they were served by the average teacher or school. In the second step, researchers compare students' actual outcomes to their predicted outcomes. The VAM for a teacher or school is the average difference—the deviation above or below the prediction—across students taught. VAMs address the following question: To what extent does the actual level of student performance exceed (or fall short of) the level that is predicted for students with similar prior achievement and background characteristics if taught by the average teacher or school?

The predicted achievement level for each student is the best estimate of how that student will do, given everything we know about him or her. We base our predictions on data from both the current and the past year—we cannot actually predict an outcome in advance for any particular student, because that would require knowing how well similar students perform in the current year. Figures I.1 and I.2 provide a simplified graphical illustration of how these predictions work.

Student achievement in prior years is the most important predictor of current student achievement. Figure I.1 draws a simple prediction line in which a student's 2013 test score is predicted based only on the student's 2012 score in the same subject. Each pair of scores (2012 and 2013) for an individual student is represented by a diamond on the chart. The line slopes up, indicating that students with higher test scores in 2012 are predicted to have higher test scores in 2013. Students with diamonds above the line performed better than would be predicted according to their 2012 test score, and students with diamonds below the line did not perform as well as would be predicted.

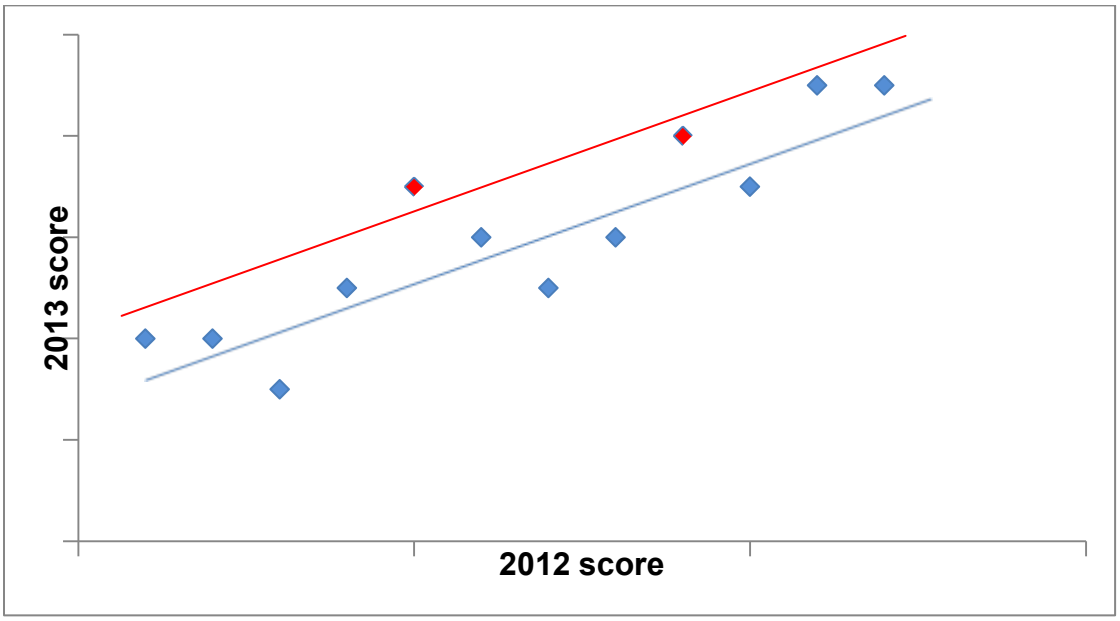
Figure I.1. Prediction based on last year's score



Note: The diamonds in this figure depict hypothetical student test score data. The line through the points represents the prediction of students' 2013 scores based upon their 2012 scores.

Test scores are not the only important student characteristics that are related to achievement. Figure I.2 shows how predictions can also take into account additional student characteristics. In this figure, the two red diamonds represent gifted students and the blue diamonds represent students who are not classified as gifted. The two gifted students have scores above the line, which suggests that gifted students, on average, did slightly better in 2013 than non-gifted students who had the same 2012 scores. By adjusting the prediction upward, we account for gifted status, meaning that we compare not only students with pre-test scores similar to each other, but also gifted students to other gifted students when making predictions. This adjusted predicted line for gifted students is represented in red.

Figure I.2. Prediction based on last year's score + gifted status



Note: The diamonds in this figure depict hypothetical student test score data. The red diamonds represent gifted students and the blue diamonds non-gifted students. The red and blue lines represent the prediction of students' 2013 scores based upon their 2012 scores for gifted and non-gifted students, respectively.

The value-added models implicitly make predictions for every student in a class or school, using data on a wide range of student characteristics from across the district or state. Combined, these predictions tell us how any particular class would do if served by the average teacher. Each teacher's value added is then measured by the average departure from prediction for all of the teacher's tested students. Teachers whose classes exceed their predicted scores are above average in value-added terms. Teachers whose classes fall short of their predicted scores have below average value added.

Value added is inherently a relative rather than absolute measure of a teacher's contribution compared to other teachers in the district or state. Calculating value added does not require assessments that are consistently scaled across grades, and a VAM result does not provide information about whether a teacher's contribution to student achievement is "good enough." The determination of the level of value added that is minimally acceptable (or that is exemplary) is a decision that must be made by policymakers, and might vary depending on the purpose for which the information is being used (thus, the value-added threshold that triggers targeted professional development might differ from the threshold used to deny tenure, for example).

The next three chapters of our report describe the student outcomes that are used to calculate Pittsburgh's VAMs (Chapter II); enumerate the information on students that is used to predict their performance and account for factors outside the control of the teacher or school (Chapter II); discuss the technical details of the VAMs (Chapter III); and explain some of the limitations of the VAMs (Chapter IV). The value-added models used in Pittsburgh are applied to a variety of exams, including the statewide Pennsylvania System of School Assessments (PSSAs) and Keystone exams, nationally normed TerraNova exams, the Preliminary SAT (PSAT), locally

developed foreign-language exams in Spanish and French, and Curriculum-Based Assessments (CBAs) in a wide range of courses at middle- and high-school levels. Models were also developed to assess student attendance and passing rates in core classes.

The last four chapters of the report explain how VAMs for each student outcome are combined to create a series of composite measures for each school and teacher (Chapter V); describe the process for locating the performance of Pittsburgh's schools in the statewide distribution of value added (Chapter VI); present summary statistics related to VAM results for Pittsburgh schools and teachers (Chapter VII); describe how the control variables included in the value-added models are related to student achievement (Chapter VIII); and discuss the application of VAMs for use in a program designed to recognize and reward outstanding performance: the Students and Teachers Achieving Results (STAR) program (Chapter IX).

II. STUDENT OUTCOMES AND BACKGROUND CHARACTERISTICS USED IN PITTSBURGH'S VALUE-ADDED MODELS

In this chapter, we describe the student outcomes and background characteristics that were used to estimate the VAMs for teachers and schools using Pittsburgh's local data. Section A pertains to test-based outcomes and baselines for prior student achievement (pre-test measures). Section B pertains to non-test outcomes and baseline measures. In Section C, we describe the other factors that were included in the value-added models to account for factors outside the control of teachers or schools, including student and peer characteristics, class size, and course type.

A. Test outcomes and baselines

We used many different assessment outcomes to calculate VAMs for teachers and schools. These outcomes can be placed into three general categories. The first is composed of state assessments, the PSSAs and Keystone exams. The second set of assessments includes other standardized tests that are administered in Pittsburgh but not given to every student in the state, the TerraNova exams and the PSAT. The third category is made up of Pittsburgh's locally developed CBAs and foreign-language multiple-choice tests (Multimode exams). Table II.1 lists each assessment along with which category it falls under. If teacher VAMs were estimated for a specific grade and assessment, a "T" appears in the corresponding cell. Similarly, an "S" indicates that an assessment was used to estimate school value-added.

Table II.1. Test scores used for VAMs by subject and grade

	Grade										
Test	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
State assessments											
PSSA Reading		S,T	S,T	S,T	S,T	S,T	S,T			S,T	
PSSA Writing				S,T			S,T			S,T	
PSSA Math		S,T	S,T	S,T	S,T	S,T	S,T			S	
PSSA Science			S,T				S,T				
Keystone Algebra I								S,T			
Keystone Literature									S,T		
Standardized tests not given across state											
TerraNova Math	S,T										
TerraNova Reading	S,T										
PSAT Reading									S	S	
PSAT Writing									S	S	
PSAT Math									S	S	

Table II.1 *(continued)*

Test	Grade										
	2	3	4	5	6	7	8	9	10	11	12
Locally developed assessments											
CBA Math					S,T	S,T	S,T				
CBA Algebra I								S,T			
CBA Algebra AB-BC								S,T			
CBA Geometry									S,T		
CBA Geometry AB-BC									S,T		
CBA Algebra II										S,T	
CBA Reading					S,T	S,T	S,T				
CBA English Language Arts								S,T	S,T	S,T	S,T
CBA African American Literature											S,T
CBA Earth Science					S,T						
CBA Life Science						S,T					
CBA Biology								S,T			
CBA Chemistry									S,T		
CBA Physics							S,T			S,T	
CBA Physical Science						S,T	S,T				
CBA Civics								S,T			
CBA Social Studies					S,T	S,T					
CBA World History									S,T		
CBA US History							S,T			S,T	
Spanish Multimode								T	T		
French Multimode									T		

Note: Cells marked with an "S" correspond to assessments that were used in the school VAMs. Cells marked with a "T" correspond to assessments that were used in the teacher VAMs. Tests that are sometimes taken out of grade by students are recorded in the grade cell where the majority of students take the test. Grade 11 PSSA exams were discontinued in the 2012–13 school year and the 8th-grade science PSSA exam was determined to no longer align with the science curriculum beginning in the 2012–13 school year. These assessments are included in this table because the teacher and school VAMs span multiple years and both of those assessments were used in earlier years of the multi-year VAM.

Most of the exams listed in Table II.1 were used in both teacher and school VAMs. There were some exceptions, however. Foreign-language VAMs were excluded from school value-added calculations because some schools have only one teacher for these subjects; reporting school-level results for individual assessments could therefore implicitly identify a single teacher's VAM. Some assessments, such as the PSAT and grade 11 PSSA math exams, were used only for schools, because PPS and the PFT have determined that those assessments are not directly aligned with specific courses, and thus student achievement on them cannot be attributed to specific teachers. The grade 8 science PSSA and Keystone biology exam were also excluded from all VAMs in 2012–13 because the exams were determined by PPS to not be well aligned

with the PPS curriculum in this year.[1] For similar reasons, PPS has chosen not to include the 11th-grade science PSSA in the school VAMs or the teacher VAMs—except for school VAMs used in PPS' Students and Teachers Achieving Results (STAR) awards, discussed in Chapter IX.

The validity of using these assessments in VAMs depends, first of all, on their validity as measures of student learning. Although no standardized assessment can provide a complete and comprehensive picture of everything a student is expected to learn, PPS assumes that the state's accountability tests (PSSAs and Keystone exams) are appropriate measures of student learning in the relevant grades and subjects. The district's homegrown CBA and foreign-language exams have not been subjected to intensive psychometric scrutiny, but they were explicitly designed by PPS to reflect the content of PPS courses. The TerraNova and PSAT exams, in contrast, were not designed to align with any particular courses, but they were developed and refined by psychometric experts.

The validity of using these assessments in VAMs also depends on the extent to which the VAMs produce results that can reliably distinguish the performance of schools and teachers. Prior research (McCaffrey et al. 2009; Schochet and Chiang 2010) has shown that estimates of a teacher's value added can have a substantial amount of imprecision if only one year of teaching is examined. We enhance the reliability of Pittsburgh's VAMs by averaging across multiple years of performance. Whenever possible, VAM estimates for schools are averaged across the last two years, and VAM estimates for teachers are averaged across the last three. Detailed results of the VAM analyses (in Chapter VII) show that almost all the VAMs were able to distinguish some schools and teachers from average.

Table II.2 shows the same list of outcome measures along with the prior test-score measures that we used as baseline controls in each VAM to account for students' own prior achievement. The grade level indicates the grade of the majority of students taking an assessment, but all students taking a particular assessment, regardless of their grade level, were eligible to be included in the VAM analysis.[2] Baseline scores were selected in consultation with PPS with two goals in mind: (1) maximize the predictive power of the model, and (2) minimize the number of students we must exclude from the model because of missing baseline test scores. Whenever possible, we used at least one baseline assessment in the same subject area as the outcome of interest. Including additional test scores, even in other subjects, improves the predictive power and precision of the model, because previous test scores in any subject provide additional information about students' baseline knowledge and abilities. Thus, all models include at least two tests from prior years. We also accounted for a third prior test in all cases in which adding the third prior test can be done without excluding substantial numbers of students (for example, those who lack an additional prior test because they transferred into PPS after the particular baseline test was taken). To minimize the number of students excluded from the model because of missing scores, we typically used only baseline tests that most students took in the year prior

[1] VAMs based on multiple years of data may still include information from the 8th-grade science PSSA, because it was aligned with the curriculum in previous years. The Keystone biology exam was not included in any value-added model because it was deemed to be not well aligned with the PPS curriculum in its single year of existence.

[2] A mismatch between the typical grade and actual grade occurs most commonly in PPS high school CBAs, which align with a particular course and not a particular grade level.

to the current test. The one exception to this rule is 8th-grade PSSA scores, which were generally available for high school students, so they were included as baseline scores in value-added models for 10th-, 11th-, and 12th-grade outcomes.[3]

Table II.2. Assessment outcomes and baseline test scores used in PPS VAMs, 2012–13

Outcome	Prior test 1	Prior test 2	Prior test 3
TerraNova Math Grade 2	TerraNova Math Grade 1	TerraNova Reading Grade 1	
TerraNova Reading Grade 2	TerraNova Reading Grade 1	TerraNova Math Grade 1	
PSSA Math Grade 3	TerraNova Math Grade 2	TerraNova Reading Grade 2	
PSSA Reading Grade 3	TerraNova Reading Grade 2	TerraNova Math Grade 2	
PSSA Math Grade 4	PSSA Math Grade 3	PSSA Reading Grade 3	
PSSA Reading Grade 4	PSSA Reading Grade 3	PSSA Math Grade 3	
PSSA Science Grade 4	PSSA Math Grade 3	PSSA Reading Grade 3	
PSSA Math Grade 5	PSSA Math Grade 4	PSSA Reading Grade 4	PSSA Science Grade 4
PSSA Reading Grade 5	PSSA Reading Grade 4	PSSA Math Grade 4	PSSA Science Grade 4
PSSA Writing Grade 5	PSSA Reading Grade 4	PSSA Math Grade 4	PSSA Science Grade 4
PSSA Math Grade 6	PSSA Math Grade 5	PSSA Reading Grade 5	PSSA Writing Grade 5
PSSA Reading Grade 6	PSSA Reading Grade 5	PSSA Writing Grade 5	PSSA Math Grade 5
CBA Math Grade 6	PSSA Math Grade 5	PSSA Reading Grade 5	PSSA Writing Grade 5
CBA Reading Grade 6	PSSA Reading Grade 5	PSSA Writing Grade 5	PSSA Math Grade 5
CBA Earth Science Grade 6	PSSA Math Grade 5	PSSA Reading Grade 5	PSSA Writing Grade 5
CBA Social Studies Grade 6	PSSA Reading Grade 5	PSSA Writing Grade 5	PSSA Math Grade 5
PSSA Math Grade 7	PSSA Math Grade 6	PSSA Reading Grade 6	
PSSA Reading Grade 7	PSSA Reading Grade 6	PSSA Math Grade 6	
CBA Math Grade 7	PSSA Math Grade 6	PSSA Reading Grade 6	
CBA Reading Grade 7	PSSA Reading Grade 6	PSSA Math Grade 6	
CBA Life Science Grade 7	CBA Earth Science Grade 6	PSSA Math Grade 6	PSSA Reading Grade 6
CBA Physical Science Grade 7	CBA Earth Science Grade 6	PSSA Math Grade 6	PSSA Reading Grade 6
CBA Social Studies Grade 7	PSSA Reading Grade 6	PSSA Math Grade 6	
PSSA Math Grade 8	PSSA Math Grade 7	PSSA Reading Grade 7	
PSSA Reading Grade 8	PSSA Reading Grade 7	PSSA Math Grade 7	
PSSA Science Grade 8	CBA Life Science Grade 7	PSSA Math Grade 7	PSSA Reading Grade 7
PSSA Writing Grade 8	PSSA Reading Grade 7	PSSA Math Grade 7	
CBA Math Grade 8	PSSA Math Grade 7	PSSA Reading Grade 7	
CBA Reading Grade 8	PSSA Reading Grade 7	PSSA Math Grade 7	

[3] We also impute missing values for these assessments. See Chapter IV, Section D for details.

Table II.2 *(continued)*

Outcome	Prior Test 1	Prior Test 2	Prior Test 3
CBA Physical Science Grade 8	CBA Life Science Grade 7	PSSA Math Grade 7	PSSA Reading Grade 7
CBA Physics Grade 8	CBA Life Science Grade 7	PSSA Math Grade 7	PSSA Reading Grade 7
CBA US History Grade 8	PSSA Reading Grade 7	PSSA Math Grade 7	
CBA Algebra I/AB-BC Grade 9	PSSA Math Grade 8	PSSA Reading Grade 8	PSSA Writing Grade 8
Keystone Algebra I Grade 9	PSSA Math Grade 8	PSSA Reading Grade 8	PSSA Writing Grade 8
CBA ELA I Grade 9	PSSA Reading Grade 8	PSSA Writing Grade 8	PSSA Math Grade 8
CBA Biology Grade 9	PSSA Science Grade 8	PSSA Math Grade 8	PSSA Reading Grade 8
CBA Civics Grade 9	CBA US History Grade 8	PSSA Reading Grade 8	PSSA Writing Grade 8
Spanish Multimode Level 1**	Spanish Multimode Grade 8	PSSA Math Grade 8	PSSA Reading Grade 8
CBA Geometry Grade 10	CBA Algebra I/AB-BC Grade 9	PSSA Math Grade 8	PSSA Reading Grade 8
CBA ELA II Grade 10	CBA ELA I Grade 9	PSSA Reading Grade 8	PSSA Writing Grade 8
Keystone Literature Grade 10	CBA ELA I Grade 9	PSSA Reading Grade 8	PSSA Writing Grade 8
CBA Chemistry Grade 10	CBA Biology Grade 9	PSSA Science Grade 8	PSSA Math Grade 8
CBA World History Grade 10	CBA Civics Grade 9	PSSA Reading Grade 8	PSSA Writing Grade 8
CBA Algebra 2 Grade 11	CBA Geometry Grade 10	PSSA Math Grade 8	PSSA Reading Grade 8
Spanish Multimode Level 2**	Spanish Multimode Level 1	PSSA Math Grade 8	PSSA Reading Grade 8
French Multimode Level 2**	French Multimode Level 1	PSSA Math Grade 8	PSSA Reading Grade 8
PSAT Math Fall Grade 10*	PSSA Math Grade 8	PSSA Reading Grade 8	PSSA Writing Grade 8
PSAT Reading Fall Grade 10*	PSSA Reading Grade 8	PSSA Writing Grade 8	PSSA Math Grade 8
PSAT Writing Fall Grade 10*	PSSA Writing Grade 8	PSSA Reading Grade 8	PSSA Math Grade 8
CBA ELA III Grade 11	CBA ELA II Grade 10	PSSA Reading Grade 8	PSSA Writing Grade 8
CBA Physics Grade 11	CBA Chemistry Grade 10	PSSA Science Grade 8	PSSA Math Grade 8
CBA US History Grade 11	CBA World History Grade 10	PSSA Reading Grade 8	PSSA Writing Grade 8
PSSA Reading Grade 11	CBA ELA II Grade 10	PSSA Reading Grade 8	PSSA Writing Grade 8
PSSA Writing Grade 11	CBA ELA II Grade 10	PSSA Writing Grade 8	PSSA Reading Grade 8
PSAT Math Fall Grade 11*	PSAT Math Fall Grade 10	PSAT Reading Fall Grade 10	PSAT Writing Fall Grade 10
PSAT Reading Fall Grade 11*	PSAT Reading Fall Grade 10	PSAT Writing Fall Grade 10	PSAT Math Fall Grade 10
PSAT Writing Fall Grade 11*	PSAT Writing Fall Grade 10	PSAT Reading Fall Grade 10	PSAT Math Fall Grade 10
PSSA Math Grade 11*	CBA Geometry/AB-BC Grade 10	PSSA Math Grade 8	PSSA Reading Grade 8
CBA ELA IV/AA Lit Grade 12	CBA ELA III Grade 11	PSSA Reading Grade 8	PSSA Writing Grade 8

Note: Grade level refers to the grade most students are in when taking the exam. PSSA-Modified exams are included and combined with the regular PSSA wherever available.

* Indicates that an outcome was included for schools but not teachers; ** Indicates that an outcome was included for teachers but not schools.

B. Non-test outcomes and baselines

Although VAMs typically involve outcomes based on tests, two of the school-level VAMs used in Pittsburgh were based on non-test student outcomes: the passage rate of core courses (in high schools) and student attendance (for grades 4 and higher). Including these non-test outcomes offers a more comprehensive view of a school's effect on students that might not be represented by test scores alone. These were not used as outcomes for teacher VAMs, because we assumed that these outcomes could not be attributed to individual teachers. The method of estimation for the non-test outcome VAMs differed slightly from the method used for test-based VAMs; see Chapter III for details.

Table II.3 lists the non-test outcomes and their baseline measures. We included baseline test measures alongside baseline measures of the outcome of interest because the test measures typically improved the predictive power of the model (that is, current attendance rate and core pass rate are related to previous achievement as well as previous attendance and core pass rates). As with the test measures, the VAMs for attendance rate and core pass rate are intended to measure the school's contribution to those outcomes, not their absolute levels. The VAMs assess whether students are doing better or worse than predicted in terms of attendance and core pass rate after accounting for student characteristics and previous performance. Attendance and core pass rate VAMs involved comparisons only within PPS, because equivalent data are not available statewide.

Table II.3. Non-assessment outcomes and baseline measures used in school VAMs, 2012–13

Outcome	Grades	Baseline measures	Baseline test, grade(s)
Attendance rate	4-8	Prior attendance rate, 3-7	PSSA math & PSSA reading, 3-7
Attendance rate	9-12	Prior attendance rate, 8-11	PSSA math & PSSA reading, 8
Core courses passed (%)	9-12	Core courses passed (%), 8-11	PSSA math & PSSA reading, 8

Note: The attendance and core pass rate models also include indicators for perfect attendance and 100 percent prior pass rate, respectively.

1. Core course pass rate

The value-added model using pass rates for core courses was designed to provide useful information on how effective a school is at moving students toward a high school diploma, accounting for their prior progress. PPS defines core courses to be those in math, reading/language arts, science, and social studies. Using Pittsburgh's course enrollment and grades data, we determined the percentage of core courses that a student passes, and then applied a value-added model to that percentage (that is, we assessed whether the percentage was better or worse than predicted, given the student's prior core pass rate and other characteristics). Ideally, we would use the number of core courses or credits that a student still needs to graduate rather than this percentage. However, the number of courses/credits needed to graduate is not measured consistently across all high schools in Pittsburgh: curriculum differences give some students greater access than others to various core courses. Using the percentage of core courses passed allows us to account for these curriculum differences.

2. Attendance rate

For grades 4 through 12, we estimated schools' contributions to students' rates of attendance during the school year, accounting for their attendance in the prior year. The measure of attendance does not distinguish between excused and unexcused absences. Although excused and unexcused absences are given separate codes in Pittsburgh's data, our examination of those data suggested that standards for determining whether an absence is excused or unexcused may vary over time and among schools. Overall absence rates were more stable over time and across schools than rates of excused or unexcused absences. We therefore used the overall attendance rates in the current VAMs. As with other VAMs, the attendance VAM did not use the raw attendance rate as the measure of school performance, but rather measured the extent to which the school's students were attending at higher or lower rates than predicted, given their attendance rates in the preceding year and other baseline characteristics.

C. Student and peer characteristics, class size, course type, and school choice

Each VAM accounted for observable student characteristics to help isolate the effect of teachers and schools on student achievement. The factors that were included in the VAMs have been found to be correlated with student performance while also being plausibly outside the control of teachers and schools. Table II.4 defines the student background characteristics that were included in teacher and school VAMs.

In addition to student characteristics, we accounted for classroom average characteristics in most models.[4] These measures of a students' peers account for classroom composition based on gender, meals program eligibility, English-language-learner status, gifted status, disability rate, prior-year absence rate, prior-year suspension rate, prior-year full-year district membership, and prior-year average PSSA math and reading scores.[5] These classroom characteristics were created by averaging individual characteristics across students enrolled in a given student's classroom. When a student takes multiple courses during the year in a subject, the classroom characteristics are averaged across the multiple classrooms.

Teacher VAMs also accounted for class size (defined using the number of students assigned to the same teacher and section of a course in a given term), which is presumably not under the control of teachers. Class size is not included as a variable in the school value-added models, however, because schools may have some influence over class size.

Pittsburgh has three main types of advanced courses at the high school level: Pittsburgh Scholars Program, Center for Advanced Studies, and Advanced Placement.[6] Because course selections are made at the beginning of the school year, they are outside the control of current-

[4] To ensure that sufficient variation was available to identify the relationship between classroom-average characteristics on test scores, we included these factors only in models incorporating multiple years of data.

[5] At the high school level, the classroom average variables for prior average PSSA math and reading scores came from grade 8 regardless of the high school year.

[6] We omit the International Baccalaureate program from the analysis because it is offered at only one Pittsburgh school.

year teachers. Thus, we accounted for students' enrollment in these programs to protect the teacher VAM estimates from being biased upward for teachers with more advanced students. We classified students as being enrolled in these programs on a yearly basis, pooling information across courses for each subject. However, we omitted the course-type variables from the school VAMs because schools may be able to influence the availability of these programs or the amount of resources designated to them.

We also received data from PPS indicating whether a parent successfully requested special services (such as attending a school outside their feeder pattern or special education services). In addition, we received a variable that indicates whether students ever entered a magnet school lottery (regardless of whether the application was successful). We included these variables in the value-added models to help control for unobserved motivation and effort levels of students and their parents that can influence achievement growth.

Not surprisingly, students' prior achievement scores have the largest relationship to their current performance. However, some student-level characteristics tend to be statistically significant as well, even accounting for other observable factors. For example, age, gender, race, meal-program eligibility, disability, and gifted status are often statistically significant predictors of achievement. Chapter VIII provides greater detail on how these variables correlate with student test scores.

Table II.4. Variables for student background characteristics in Pittsburgh teacher and school VAMs, 2012–13

Background variable	Definition
Male	Male gender
Meals program	Free or reduced-price meal eligibility status
Race/ethnicity	African American, white, Asian, Hispanic, other race
English-language learner	English-language-learner status
Gifted	Participation in the gifted program
Pittsburgh Scholars Program*	Taking a class in the Pittsburgh Scholars Program
Advanced Placement*	Taking an advanced-placement class
Center for Advanced Studies*	Taking a Center for Advanced Studies class
Specific learning disability	SLD designation under Individuals with Disabilities Education Act (IDEA)
Speech or language impairment	SLI designation under IDEA
Emotional disturbance	ED designation under IDEA
Intellectual disability	ID designation under IDEA
Autism	AUT designation under IDEA
Physical/sensory impairment	An IDEA designation for hearing impairment, visual impairment, deafness-blindness, or orthopedic impairment
Other impairment	An IDEA designation for other health impairment, multiple disabilities, developmental delay, or traumatic brain injury
Mobility	Transferred schools during prior school year
Grade repeater	Repetition of the current grade

Table II.4 *(continued)*

Background variable	Definition
PSSA-Modified	Student took modified version of the PSSA
AB-BC Curriculum-Based Assessment	Student took the AB-BC version of the Curriculum-Based Assessment
Absence rate (prior year)	Prior-year absences divided by days of enrollment. This variable is top-coded so that its maximum is 0.50.
Suspension rate (prior year)	Prior year days suspended out-of-school or expelled divided by days of enrollment. This variable is top-coded so that its maximum is 0.20.
Full-year district membership (prior year)	Enrolled the entire prior school year in Pittsburgh
Magnet applicant	Has ever applied for entry to a magnet program
Age	Student age in years as of the beginning of the academic year, including fractional years
Behind grade for age	Student age is one or more years older than typical for grade level
Special services	Ever applied for special services, such as attending a school outside of feeder pattern or special education services

*Indicates that the variable is excluded from school value-added models. All variables are binary except absence rate, suspension rate, and age. We aggregated several of the lowest-incidence disabilities because some individual categories do not contain even a single student at all grade levels.

This page has been left blank for double-sided copying.

III. TECHNICAL DETAILS OF PPS VALUE-ADDED MODELS

This chapter explores the technical details of the value-added models estimated for PPS. We first detail the models used and then discuss how we prepared our data for analysis and processed the output to calculate the final VAMs reported to teachers and schools. A reader uninterested in the technical details of the value-added model estimation may scan or skip this chapter without loss of understanding.

A. Detailed Value-Added Model description

The following general statistical equation describes the value-added models:

$$Y_{i,t,c} = B_{i,t-1}\alpha + X_{i,t}\gamma + \bar{X}_{i,t,c}\theta + D_{i,t}\delta + T_{i,t}\tau + e_{i,t,c}. \quad (1)$$

In the model, $Y_{i,t,c}$ is the outcome for student i in year t and classroom c. The models were estimated separately for each grade, subject, and assessment. For example, $Y_{i,t,c}$ could be a student's score on the grade 5 math PSSA during 2012–13. $B_{i,t-1}$ is a vector of baseline scores for student i from a prior year to account for students' own academic histories. The baseline scores typically came from the previous school year, though baseline scores could come from up to three prior years at the high school level because 8th-grade PSSAs were used as control variables in all high school VAMs. We included two or three baseline scores rather than one, because each additional assessment improves the predictive power of the model. Whenever possible, at least one baseline score came from the same subject area as the outcome measure.[7] See Table II.2 for the full list of outcome measures and baseline variables.

$X_{i,t}$ is a set of variables for observable student characteristics, while $\bar{X}_{i,t,c}$ represents observable peer characteristics (listed in Table II.4 and described in Section C of Chapter II).[8] $D_{i,t}$ is a set of variables for a student's teachers in the subject of interest or schools during the year, $T_{i,t}$ is a set of indicators for different years, and $e_{i,t,c}$ is the error term. The coefficients in α, γ, θ, and τ are the estimated relationships between student outcomes and each respective variable, accounting for the other factors in the model. The δ symbol refers to a set of coefficients as well, one for each teacher or school in the value-added model. Each δ coefficient identifies a teacher's contribution or a school's contribution to student learning—the extent to which the actual achievement of students tends to be above or below what is expected for the average teacher or school.

We defined the average VAM score (that is, the average δ coefficient) to be zero, but this does not mean that student learning is zero for the teacher or school with the average VAM

[7] An example of an exception is the grade 4 science PSSA, where a same-subject baseline score was not available. Both baseline scores come from other subjects in these cases (for example, grade 3 math and reading PSSAs).

[8] Some of the $X_{i,t}$ and $\bar{X}_{i,t,c}$ variables are correlated with each other. Including related variables in VAMs does not mean that teacher or school effects will be estimated inconsistently. In fact, it typically improves the validity of VAM estimates so long as both the related variables are relevant to student achievement growth.

Note that we identified the relationship between student and classroom characteristics using variation within teachers and schools.

score. Rather, it means that positive VAM estimates represent above-average (above predicted) teacher or school performance, and negative VAM estimates represent below-average (below predicted) teacher or school performance. For school-level test-based VAMs, average performance was defined at the state level by creating a hypothetical distribution of statewide performance in a separate step, as described in Chapter VI. For school-level attendance and core course pass rate VAMs, and all teacher VAMs, average performance was defined among Pittsburgh schools and Pittsburgh teachers.

The VAMs for teachers and for schools differ in the number of cohorts of student data they include. Value-added models for schools examine two years of teaching, producing an average of a school's performance across the 2011–12 and 2012–13 school years. Models for teachers examine up to three years of teaching, producing an average of the teacher's performance across the 2010–11, 2011–12, and 2012–13 school years. The VAMs for teachers who have not taught in tested grades and subjects in all three years are based on the years they taught the relevant courses.

We utilized multi-year VAM estimates when possible because, compared to single-year estimates, they are less prone to random fluctuations that stem from a teacher being assigned a few students who display unusually high or low achievement due to chance.[9] VAMs based on multiple years of data can therefore detect performance differences with greater reliability, which is advantageous for any high-stakes application (see McCaffrey et al. 2009). However, multi-cohort VAMs are less reflective of immediate past performance, because they average annual value-added scores over multiple years.

B. Standardization

Because VAM estimates reported in assessment units (for example, PSSA scaled-score points) are not necessarily comparable across tests, grades, subjects, or years, we standardized all outcome measures prior to running the analyses. Specifically, we mapped assessment units to a standard measure, called a z-score, by subtracting the average value (for example, the average grade 4 math PSSA scaled score) from individual scores by school year and then dividing by the standard deviation (SD) of scores. Expressing scores like this allows us to interpret above-average scores in terms of how close to average most students tended to fall, regardless of the assessment. When estimating the value-added models, we mean-centered the data for all baseline/background variables based on the analysis sample for each VAM and exclude the constant term. This latter standardization process and the exclusion of the constant term means that the teacher and school effects were estimated relative to the contribution of the average teacher or school in the sample.

C. Teacher and school dosage

When a student was in a teacher's class for only part of the school year, we used a dosage approach to account for the amount of time that teacher had to influence the achievement growth of this student compared to that of students in the teacher's class for the full year. Dosage can be

[9] Information on an exploratory analysis from an earlier school year that compared the precision of single-year VAM estimates in Pittsburgh with that of three-year VAM estimates can be found in the appendix of Johnson et al. 2012.

thought of as a weight that is applied to each student when calculating a teacher's value-added estimate. The dosage each student has with a teacher is equal to the fraction of the school year that the student was in the teacher's class. For example, if a student moves schools in Pittsburgh during the year and was enrolled in one math class while at each school, the teacher and school dosage values in the VAM are fractions between 0 and 1 based on the days enrolled at each school.[10] We split teacher dosage values evenly across teachers when students take multiple courses in the same subject at the same school. We used the "Full Roster Method" to account for differences in dosage (Hock and Isenberg 2012; Isenberg and Walsh 2013). In 2012–13 and past years, only one teacher of record was responsible for each class. In future years, however, PPS may provide data on multiple teachers of record for each class. Using the Full Roster Method, we will be able to produce VAM estimates for teachers in co-teaching arrangements under the assumption that each teacher receives equal credit for the students they co-teach.

To determine which teachers should receive value-added estimates we used course attribution data obtained from PPS that ties courses to specific assessments used for VAMs. Only teachers who teach a VAM-attributed course in a given subject are eligible to receive a value-added estimate for that subject. For example, only teachers who teach math courses explicitly tied to the CBA geometry grade 10 exam will receive VAMs for this subject. Teachers who teach other 10th-grade math courses will not receive CBA geometry estimates (even if their students happen to take the CBA geometry exam).

D. Shrinkage

Value-added models typically use a procedure known as empirical Bayes estimation or shrinkage to address the fact that among teachers/schools with the same level of true performance, those with fewer students in the estimation sample face a greater likelihood that their students happen, by chance, to have atypically high or low learning growth driven by other factors. In the absence of a shrinkage adjustment, teachers with fewer students—that is, those with less precise estimates—will tend to be overrepresented at both the high and low ends of the estimated performance distribution just by chance.

We used an empirical Bayes procedure based on Morris (1983) to shrink our teacher and school VAM estimates.[11] The shrinkage adjustment accounts for the fact that an estimate with greater precision carries greater strength of information about a teacher's true performance level. The adjusted estimate is a weighted average of the individual's initial estimate and the mean estimate across teachers, with more precise initial estimates receiving greater weight. In essence, teachers and schools are assumed to be average in performance until evidence justifies a different conclusion. We use shrinkage for all reported estimates.[12]

[10] Students who spend nine or fewer days enrolled in a school were assigned a dosage of zero.

[11] The procedure reduces the mean squared error of the value-added estimates.

[12] For teacher VAMs, the shrinkage adjustment was applied to the assessment-level VAMs and the composite teacher VAM is a weighted average of the assessment-level VAM estimates. Chapter V contains more details about the construction of VAM composites.

To further minimize the risk of making erroneous conclusions on the basis of imprecise estimates, we reported the VAMs only of teachers who taught more than 10 students during the year.[13] This type of restriction, common in the research literature, reduces the potential for teacher effects to be influenced just by the scores of one or two students (Kane and Staiger 2002; McCaffrey et al. 2009). We used this same process in school models as well, but the restriction was generally not binding because most schools have many more than 10 students taking each assessment.

E. Accounting for measurement error

Test scores are imperfect measures of student ability. To account for this imprecision, we used a specification that includes an errors-in-variables measurement error correction (Buonaccorsi 2010). The errors-in-variables correction helps to alleviate problems related to measurement error in pretests by incorporating into the VAM information about the reliability of those tests from their publishers.[14] To implement this procedure we used grade- and subject-specific test reliability data available from test publishers' websites.[15] Almost all the VAMs accounted for prior achievement on assessments for which reliability information is available (PSSA or PSAT exams). In some VAMs, we also included controls for prior-year achievement on CBAs, TerraNova, or Multimode exams. Reliability information was unavailable for these assessments. To account for measurement error in these instances, we used the reliability of the PSSA in the same subject and year and in the nearest grade. While using the reliability information of another test is not ideal, it is preferable to assuming that these exams are perfectly reliable. The measurement error correction was implemented using the two-step estimation procedure described in Isenberg and Hock (2012).

F. Normal Curve Equivalent reporting units

Each VAM estimate is reported to teachers or schools as a normal curve equivalent (NCE) ranking.[16] NCEs are similar to percentile ranks in that they are on a 100-point scale with an average of 50. An NCE is equivalent to a percentile rank at scores of 1, 50, and 99. However, unlike percentiles, the NCE scale is an equal interval scale, which means that a given difference in NCEs represents the same difference at any point on the scale. For example, the difference in value-added between an NCE of 60 and 70 is the same as the difference in value-added between an NCE of 80 and 90 (which is not true of percentile scores).

[13] To account for the fact that some students transfer out of a teacher's class mid-year, each student was weighted by their dosage in calculating whether a teacher taught more than 10 students.

[14] Reliability is the fraction of the variance in scores that represents true differences in performance rather than random measurement errors. Random measurement errors can stem from a variety of factors, such as the health of the student on the testing day, the set of questions in the assessment, or any disruptions that might occur while students are taking the exam.

[15] PSSA reliability data can be obtained from technical reports on the Pennsylvania Department of Education website: [www.portal.state.pa.us/portal/server.pt/community/technical_analysis/7447]. PSAT reliability data can be obtained from the College Board at [www.collegeboard.com/prod_downloads/counselors/psat/PSATscores.pdf].

[16] The exception to this reporting convention is a school's STAR ranking, which is reported as a percentile rank.

For teachers, the NCE rank is an estimate of where they stand in the distribution of teachers teaching the same subjects and grades *within PPS*. For schools, in contrast, we report an NCE that estimates where they stand in the distribution of schools serving the same grades *across Pennsylvania*. Ideally, we would use a statewide comparison for all estimates, but many teachers have VAMs based on student assessments conducted only in Pittsburgh, precluding a statewide comparison.

G. Adjustment factor for Pittsburgh Westinghouse teachers

Many schedule disruptions occurred at one school, Pittsburgh Westinghouse, during the first half of the 2011–12 school year, when the school transitioned from being a grade 9–12 high school operating on the traditional semester system to being a grade 6–12 academy operating on a trimester system. PPS believes that these schedule disruptions were severe enough that teachers at Pittsburgh Westinghouse did not have the same opportunity as other PPS teachers to contribute a year's worth of learning to their students' academic growth and that teacher VAM estimates would thus not accurately represent the true ability of these teachers to contribute to student achievement growth. Note that PPS believes that both test scores and learning were systematically lower at Pittsburgh Westinghouse during the transition year. That is, the drop in test scores for Pittsburgh Westinghouse students represents a real decrease in achievement and not simply an issue with testing.

For a more accurate estimate of the contribution of Westinghouse teachers to student achievement growth, the amount of any decline in Westinghouse's *schoolwide* value-added that occurred in 2011–12 was added back to the value-added estimates for Westinghouse teachers in 2011–12.[17] The 2011–12 teacher VAMs for Pittsburgh Westinghouse teachers thus were adjusted based on the difference between the one-year 2011–12 Pittsburgh Westinghouse test-based composite school VAM estimate and the average of the one-year school VAM estimates from Westinghouse in the prior two school years.[18]

This adjustment required two main assumptions: (1) changes in school value-added estimates accurately represent changes in average teacher value-added estimates; and (2) any changes in the average estimated teacher value-added scores at Westinghouse from the two prior years and the scores for 2011–12 were due to schedule disruptions rather than true changes in teacher performance. The first assumption is reasonable because we accounted for a very similar set of student and classroom characteristics in the school and teacher value-added models and use a similar set of assessments. The second assumption was made because PPS determined that the schedule disruptions were outside teachers' control and that teacher value-added estimates should not decline as a result of them. While it is not possible to statistically disentangle any decline in overall teacher scores from declines attributable to the schedule disruptions, the adjustment was made in the interest of fairness to prevent the disruptions from negatively impacting teacher VAM scores.

[17] The scores on the teacher VAM reports include the adjustment factor for Westinghouse teachers.

[18] We used the average of school VAM estimates from the two prior years rather than focusing only on the 2010–11 estimate to increase the precision of the historical effectiveness estimate.

To adjust teachers' value-added estimates, we first averaged the Westinghouse one-year test-based composite NCE VAM scores from 2009–10 and 2010–11. We then subtracted from this average the corresponding VAM score from 2011–12, resulting in a difference of 23 NCE units. This factor was then divided by the number of years of teaching included for that teacher to ensure that the adjustment was not applied to teaching data in years before or after 2011–12.[19] VAM scores for teachers who served at Westinghouse in 2011–12 will continue to receive the adjustment through 2013–14 because these scores are based on three years of data and would thus continue to be influenced in those years by the schedule disruptions of 2011–12. After 2013–14 year, the adjustment will no longer be made.

H. Re-estimation of VAMs for teachers with no new students in a given year

It is possible that a teacher's estimated VAM will change from year to year even if the teacher has taught no new students. This occurs because the set of students taking relevant assessments during the most recent year has been added to this year's models, and student achievement from four years ago has been omitted. This difference in the sample of students entering the model results in changes to the estimated coefficients on all variables in the model, including the teacher VAMs. To avoid introducing variation in the VAMs over time unrelated to actual changes in teaching efficacy, PPS has decided not to re-estimate VAMs for teachers with the exact same set of students included in the value-added models as the preceding year. Thus, teachers contributing the same students to a value-added model in two consecutive years will receive the same VAM.

I. Technical details for VAMs for non-test outcomes

As core pass rate and attendance have maximum values attained by a sizable number of students (that is, perfect attendance, 100 percent pass rate), their VAM specifications must differ slightly from the primary model that is described by Equation (1) at the beginning of the chapter. For both these VAMs, we used a Tobit version of Equation (1). The Tobit model separately estimates the probability that an outcome will be at the ceiling of the distribution and accounts for this probability when calculating the coefficient estimates (Tobin 1958).

Core Pass Rate: About 75 percent of PPS high school students pass all their core classes each year, which means that they reached the upper limit of the core pass rate metric. In situations like this, ordinary linear regression models like Equation (1) provide biased estimates, because the relationship between higher values for the background variables and a higher core course pass rate becomes nonlinear when students pass all courses. To account for this bias, we used a Tobit model to estimate the VAM. In the core pass rate VAMs, we accounted for a student's prior year core pass rate and grade 8 PSSA math and reading scores. To allow the effect of prior year core pass rates to be nonlinear for students who

[19] For example, if a teacher taught students in all three years covered by the VAM, the adjustment factor was divided by three before being added to the teacher's value-added estimate. If a teacher had multiple years of teaching data, the adjustment was made only to VAMs tied to classes the teacher taught at Westinghouse in 2011–12. For example, if a teacher taught chemistry classes in 2010–11 and biology classes in 2011–12 at Westinghouse, only the biology VAM score was adjusted. The adjustment to the composite VAM score was weighted by the fraction of courses taught at Westinghouse in 2011–12 relative to the total number of VAM-eligible courses taught in the prior three years.

passed all core courses in the prior year, we included an indicator variable equaling one if the student had a perfect pass rate in the prior year.

Attendance Rate: Because about 5 percent of students contributing to the attendance VAM have perfect attendance each year, we estimated this VAM using the Tobit model as well. We included an indicator for perfect attendance in the prior year along with the baseline variables for a student's prior year attendance rate, PSSA math score, and PSSA reading score. When we attempted to estimate attendance rate value-added models using data before grade 4, we found that the results were very imprecise as a result of relatively small variation in attendance rates among students in these grades. We therefore limited the attendance VAMs to grades 4 and higher.

This page has been left blank for double-sided copying.

IV. METHODOLOGICAL LIMITATIONS

In this chapter, we describe some key limitations of the VAMs used in Pittsburgh.

A. Non-random assignment of students

Students are not necessarily randomly assigned to teachers and schools. If this issue is not addressed adequately, VAM estimates can be biased (Rothstein 2010). Goldhaber and Chaplin (2012) find, however, that the sorting bias is small relative to the variability of teacher value-added scores. The value-added models used for PPS assume that assignment is as good as random once we have accounted for a large set of observable characteristics. Even though this assumption may not be strictly true, research suggests that resulting bias in the VAM estimates is likely to be small. Kane and Staiger (2008) and Kane et al. (2013) found that variation in teacher VAM scores significantly predicted achievement differences in a subsequent year when classrooms were assigned randomly. Although this indicates bias is small in a particular case (when a principal is willing to randomly assign students to teachers), work by Chetty et al. (2013) suggests the finding holds even when assignment is nonrandom. The authors analyzed the average VAM of teachers at schools before and after the exit of high value-added teachers. If these teachers had high VAMs simply because they taught more talented students, one would expect the value-added of teachers remaining in the schools to rise after the transfer. Instead, Chetty et al. find little movement in the VAMs of teachers remaining in the school. This suggests that a teacher's VAM will reflect his or her ability, and not some underlying characteristic of the students he or she teaches. Together, these studies indicate that any bias that may exist does not prevent VAMs from identifying an important component of school and teacher performance.

B. Distinguishing between school and teacher effects

When we estimate teacher contributions to student learning through value-added models, some of the effect we attribute to a teacher may actually be due to the school where the teacher works. This could occur, for example, if the school provides a better working environment, or if it gives teachers more preparation time as compared to other schools. It is possible to include school indicators to account for the influence a school may have on a teacher's effectiveness. However, including school indicators means that teachers will be compared only within the same school rather than across the district. This would create an undesirable "zero-sum game" within schools, in which teachers can raise their value-added only by doing better than their colleagues down the hall. It would also be likely to underestimate true teacher effects, because taking out the average performance in the school is likely to remove some of the teacher-specific performance as well. To avoid these problems, we did not include school indicators in the teacher VAMs.

An alternative method to account for the influences of schools in teacher-level VAMs would be to add variables accounting for school characteristics. We cannot adjust for most school characteristics that might be directly relevant to teacher value-added (for example, resources available, principal quality, school safety), because data are not readily available on those characteristics. Even if these data were readily available, it is difficult to separate the effect of a school having good characteristics from the possibility that good teachers choose to work at schools with attractive characteristics; doing so requires variation in characteristics for the same school over time and substantial transfer of teachers across schools. Nonetheless, we performed

exploratory analyses that included school-level measures of characteristics that could affect teacher value-added and are available in Pittsburgh's data, such as school-wide averages of the number of days students are suspended, percentage of students eligible for free or reduced-price meals, and prior student test scores. The exploratory analyses found that these school characteristics explained very little of the variation in student outcomes. More to the point, the teacher value-added estimates were almost identical with or without the inclusion of the school characteristics. The lack of explanatory power of these variables could occur because these factors are not strongly related to the actual factors influencing teacher effectiveness or because these characteristics vary little within schools over time. If school-level data that are more directly relevant to teacher value-added become available in the future, we could examine the possibility of including such data, but for now we omit school variables from the teacher VAMs.

C. School VAMs do not use "Pre-treatment" baselines

Teacher and school value-added models are nearly identical in their analytic structure—differing only in whether teacher or school dosage variables are used and with respect to a few control variables—but there is a substantive difference between the models related to the baseline scores. Specifically, the baseline scores used for most grades in the school models are not "pre-treatment" measures of student achievement as they are for teachers. Except in entry grades (for example, 6th grade in a school with grades 6 to 12), students are generally served by the same school both in the current year (the year to which a set of VAM estimates apply) and in the prior year, when baseline scores are measured. This implies that some variables that we assume are outside the control of a school for a current year value-added model were actually affected by that school in the prior year. For example, a school could hold a student back for a grade, which would affect next year's VAMs differently from the way it would if the school had allowed the student to progress to the next grade.[20]

We could instead use school-level VAMs in which baseline scores are always measured before the student entered the current school. But this has the great disadvantage of excluding large numbers of students from the analysis, especially in K-5 and K-8 schools, since no pre-kindergarten measure of achievement exists. We therefore conducted a sensitivity analysis that examined whether using last year's score produced results similar to using pre-entry baseline scores at the middle and high school levels. Results were very similar. Because baselines from last year produce results that are similar to those produced by "pre-entry" baselines, and because we do not want to remove large numbers of students from the analyses, our models typically rely on baseline scores from last year for school VAM estimates as well as teacher VAM estimates.

D. Missing and omitted data

Value-added models can account only for factors that are measured in the data, which means that estimates may be biased if important background/baseline variables cannot be included.

[20] This is not typically an issue in teacher models, because students generally change teachers each year (except in instances when teachers "loop" to the next grade with their students, in which case the VAM operates like a school-level VAM).

1. Missing data on resources for students and schools

Pittsburgh's data system does not currently track student participation in all programs (during the school year or in summer school) that may help raise test scores. Participation in some programs, including a number of after-school activities and the Summer Dreamers Academy (a free educational summer camp open to all PPS students in kindergarten through 7th grade), is tracked but was not included in the statistical models. We examined the suitability of the inclusion of control variables for these programs in value-added models in an earlier analysis (Rotz et al. 2013). We determined that it is possible for teachers to play a role in student enrollment in many after-school programs. Thus, controlling for students' enrollment in after-school programs could actually lead to greater bias in VAMs, because students' participation is attributable in part to the influence of a teacher. Participation in the Summer Dreamers Academy occurs in the summer before a student is taught by a teacher, so it would not be problematic to include a control variable for this program. However, when we experimented with adding a control for Summer Dreamers Academy enrollment, we found that the teacher VAM estimates were not sensitive to the inclusion of this variable. Thus, rather than making an exception for Summer Dreamers, we recommended that, for simplicity, our statistical model not account for any extracurricular activities.

1. We also lack information on the number of hours of instruction in particular subjects, meaning that we cannot account for some teachers or schools spending more time, for example, on social studies than do other teachers or schools. Ignoring differences in instructional hours or available resources could be problematic to the extent that they are outside the control of a teacher or school.

2. Missing baseline test scores

In each value-added model some students who had data on the outcome measure were dropped because of missing data on at least one baseline/background variable. In most cases, the missing element was a prior test score.[21] Prior scores could have been missing for several reasons, such as if students transferred into Pittsburgh from outside the district, took a test out of grade, or were absent from school during testing in the prior year. To increase precision, we could impute the missing prior test scores for these students and thus keep them in the model. Imputation involves using data on other previous test scores to estimate a value for the missing prior-year score that is used as a baseline in the model. However, it can be difficult to find a previous test score that can be used to impute the missing values consistently for each model. Also, imputation is not feasible for students who transferred to Pittsburgh from other districts, because there is no information in Pittsburgh's data collection on any of their prior scores. Therefore (with the exception described in Section IV.D.3), we did not impute missing values.

Missing data tend not to be a persistent problem for most students over time. For example, if a student transferred from another district in 2011–12 and was missing prior test score data, he or she would have taken the normal end-of-year assessments in Pittsburgh. That student would have been dropped from the 2011–12 VAMs due to missing prior test score data, but would have appeared in the 2012–13 VAMs, because the end-of-year assessments in 2011–12 could be used

[21] Across all teacher value-added models, the median percentage of students excluded from the sample due to missing data was 13 percent.

as the baseline scores needed in the 2012–13 VAMs. Therefore, except in cases of rapid mobility, students tended to be picked up by the VAMs after a full year in the district.

3. Substituting 9th-grade entry SRI scores for missing 8th-grade PSSA scores

Although missing baseline scores are not problematic in most cases, high-school entry is a special case. In high school value-added models, we used a prior-year score in the same subject as the primary baseline score. We also included the 8th-grade PSSA score that is in the most closely related subject to the outcome variable as the additional baseline score.[22] Thus, a student missing 8th-grade PSSA scores would be permanently excluded from all the high school VAMs. This can be problematic for students who in 9th grade transfer into Pittsburgh high schools from private or parochial middle schools and thus lack prior PSSA scores.

To prevent the exclusion of a relatively large fraction of students from the high school VAMs, we imputed values for missing 8th-grade PSSA scores using data on student achievement at the beginning of 9th grade. On average, the imputation increases our VAM sample sizes by about 10 percent. Starting in 2010–11, the Scholastic Reading Inventory (SRI) was given to all PPS students in September of 9th grade. We used the current year's 9th-grade SRI to impute prior-year values for missing 8th-grade PSSA reading, writing, math, and science scores.[23] The 9th-grade SRI score gives us a way to estimate what the student's 8th-grade PSSA score would have been.[24] Performing this imputation may be important to avoid bias in the results, because the number of 9th-grade students who are missing 8th-grade PSSA scores varies widely across Pittsburgh's high schools.

4. Missing data on classroom average characteristics

In some cases, we had data available on a student but not his or her peers. This typically happened when we could not assign a student to a particular classroom. Because we included controls for classroom average characteristics in our value-added models, these students will be omitted from our analysis unless their characteristics are imputed. We used the simplest imputation possible for these classroom-average characteristics: setting any missing value to the average across all PPS students without these data missing. In essence, this assumes that unassigned students have an average peer group. By making this assumption, we are able to boost our sample size without adding additional complexity to the model estimation.

[22] The 8th-grade PSSA was used as an additional control rather than using another prior year CBA in a different subject in order to maximize sample size. Because many high school students take courses in different orders, students often did not take both the prior CBAs we would otherwise have used as control variables.

[23] Transfers from out of district are often missing prior-year attendance and suspension data that we account for in the VAMs, so we impute these values as well. We used a single imputation method based on the conditional distribution of 9th-grade SRI scores and other student characteristics. See Schafer and Graham (2002) for details and information on the statistical properties of this method.

[24] We assumed that high schools have not yet had a chance to influence student achievement before the administration of the 9th-grade SRI exam.

E. Floor and ceiling effects

Some educators may be concerned that having students who achieved a perfect score on a baseline assessment could lead to unfairly low estimates of value-added. The concern is based on the premise that because value-added measures growth over the course of the year, a student who scored perfectly on the baseline assessment would have no potential for measurable growth (a ceiling effect). A similar concern exists for students who received the minimum test score at baseline. These students could potentially have "nowhere to go but up" (a floor effect).

It is unlikely that these effects led to substantial biases in the PPS context, for two main reasons. First, the value-added model does not literally measure changes in student test scores. Instead, it uses a linear function of past test scores and other student characteristics to predict current test scores. The difference between that prediction and actual scores is the teacher's or school's value-added. Even when students score perfectly on all past tests, they are generally not predicted to score perfectly on the current year's exams. Thus, even students who start with scores at the ceiling can positively impact teacher value-added. A similar argument holds for students starting with the minimum possible score on a test. These students are generally predicted to score above the minimum on the subsequent year's test, which means that they would need to perform even better than this prediction to positively impact teacher value added (Resch and Isenberg 2014).

In addition, few PPS students score at the floor or ceiling. Koedel and Betts (2010) and Resch and Isenberg (2014) have shown that ceiling effects are really only a concern when a large proportion of students have scores at the maximum value. In the assessments used by PPS in the 2012–13 school year, about 0.6 percent of student test scores were at the maximum and 0.7 percent were at the minimum. Table IV.1 further breaks down the students who scored at the maximum or minimum value, by test and subject, pooling across grades. For almost all assessments, maximum and minimum scores are very rare. At most, some assessments, such as the TerraNova and Multimode tests, have between 2 and 3 percent of students scoring at the maximum. Further, minimum scores are more common on the PSSA Science exams (particularly for 4th grade). However, these rates are still quite low and thus suggest that floor and ceiling effects are not a substantial issue for the PPS VAMs.

Table IV.1. Share of students scoring at maximum value, by test type and subject, 2012–13

Test	Students taking assessment	Students scoring at minimum value		Students scoring at maximum value	
		Number	Share	Number	Share
TerraNova Math	3640	2	0.001	90	0.025
TerraNova Reading	3631	2	0.001	71	0.020
PSSA Reading	10609	70	0.007	10	0.001
PSSA Writing	3483	47	0.013	11	0.003
PSSA Math	10689	8	0.001	22	0.002
PSSA Science	1736	137	0.079	2	0.001

Table IV.1 *(continued)*

Test	Students taking assessment	Students scoring at minimum value		Students scoring at maximum value	
		Number	Share	Number	Share
CBA Math Exams	8569	5	0.001	24	0.003
CBA ELA Exams	9303	1	<0.001	41	0.004
CBA Science Exams	8610	9	0.001	6	0.001
CBA Social Studies Exams	8121	3	<0.001	12	0.001
Spanish Multimode	1573	1	0.001	43	0.027
French Multimode	513	0	0.000	10	0.019
Keystone Algebra I	3601	1	<0.001	1	<0.001
Keystone Literature	2211	1	<0.001	1	<0.001
PSAT Reading	2754	2	0.001	8	0.003
PSAT Writing	2754	53	0.019	0	0.000
PSAT Math	2754	8	0.003	2	0.001

Source: Authors' calculations based on data provided by PPS.

F. Adverse incentives of using VAMs to reward teachers and schools

Value-added models are valuable tools for identifying and rewarding effective teachers and schools, but rewarding educators based on certain outcomes can have potentially adverse incentive effects. Teachers may respond to bonuses associated with their VAMs by "teaching to the test" or otherwise deviating from behavior that would maximize student learning more broadly. In extreme cases, the use of VAMs for evaluation purposes might induce some teachers to cheat on the exams. The use of multiple measures in addition to VAMs to evaluate teachers (principal observations and student surveys) should mitigate this concern somewhat, because most of a teacher's evaluation is based on measures other VAMs. Nonetheless, it is important that PPS maintain strong test security procedures to protect the integrity of the exams.

VAMs based on non-test outcomes (such as our core pass and attendance rate VAMs) are also susceptible to incentive problems. Rewarding administrators based on the grades and attendance rates of their students could lead educators to be more lenient when choosing whether to fail students or mark them as absent. It will therefore be important that PPS ensure that the standards for passing a core course or being marked absent remain constant as these non-test outcomes continue to be used for school evaluation.

V. COMPOSITE VALUE-ADDED MEASURES

Every school in Pittsburgh has VAMs based on several different outcomes, with assessments spanning multiple grades and covering a variety of subjects. Many teachers likewise have VAM estimates related to more than one student assessment. At the policy direction of PPS and the PFT, the VAM results for the individual test-based outcomes were aggregated into composite measures for reporting purposes and for informing awards for STAR schools (described in Chapter VIII). This creates a simpler presentation of results and allows educators to get a sense of a school's subject-wide and overall performance. Composite measures were calculated separately by subject (for schools only) and across all test-based measures (for schools and teachers). For example, the math composite for a middle school incorporates VAM data from math PSSAs and CBAs in grades 6, 7, and 8.

A. Composition of composites

Table V.1 shows how all the individual assessments are grouped into subject-wide composites for Pittsburgh's school VAMs The composite across all test-based measures for an elementary school, for example, includes all TerraNova exams in grade 2 and PSSAs in grades 3, 4, and 5. Composites were calculated based on the assessments that were available in the grade ranges taught at a school. Schools with grade configurations of K-8 or 6-12 received composite scores that included all assessments from the relevant grades. Pittsburgh schools were then ranked all together based on their effectiveness rating for each composite. (STAR awards use a different composite, described in Chapter VIII).

Table V.1. The composition of subject composites for Pittsburgh school VAMs, 2011–13

	Elementary school grades K to 5	Middle school grades 6 to 8	High school grades 9 to 12
Math composite	TerraNova math (2) PSSA math (3,4,5)	PSSA math (6,7,8) CBA math (6,7,8)	CBA algebra I/AB-BC (9) Keystone algebra I (9) CBA geometry (10) PSAT math (10, 11) PSSA math (11) CBA algebra II (11)
English/language arts composite	TerraNova reading (2) PSSA reading (3,4,5) PSSA writing (5)	PSSA reading (6,7,8) CBA ELA (6,7,8) PSSA writing (8)	CBA English I (9) CBA English II (10) Keystone Literature (10) PSAT reading (10, 11) PSAT writing (10, 11) PSSA reading (11) PSSA writing (11) CBA ELA III (11) CBA ELA IV/AA literature (12)
Science composite	PSSA science (4)	CBA earth science (6) CBA life science (7) CBA physical science (7,8) PSSA science (8) CBA physics (8)	CBA biology (9) CBA chemistry (10) CBA physics (11)
Social studies composite	n/a	CBA social studies (6,7) CBA US history (8)	CBA civics (9) CBA world history (10) CBA U.S. history (11)

Note: The grade level of the majority of students follows each assessment in parentheses.

Fourth-grade PSSA science and middle school social studies CBAs were not reported in subject-specific composites for K-5 and 6-8 schools because some schools have only one teacher for these subjects; reporting results for individual assessments would therefore implicitly identify a teacher. They were, however, included in subject composites for K-8 and 6-12 schools, where they are combined with other VAMs (and other teachers). They were also included in the overall test-based composite (averaged across subjects) for each school in all relevant grade configurations.

Foreign-language VAMs were similarly excluded from school value-added models. Even if these measures were combined into a foreign-language composite at the school level, some schools may have only one Spanish or French teacher, which means that the information could be used to identify an individual teacher's value added. In addition, as is described in more detail in Chapter VI, the school composites rely on a mapping between state-level and district-level VAMs. No state-level foreign-language assessments exist, which means that the mapping would be possible only under strong assumptions about the similarity between school effectiveness in foreign-language instruction relative to other subjects.

Because many teachers receive VAM estimates based on only one or two assessments in the same subject, subject-level composites were not reported for teachers. Instead, the VAM score for each assessment was reported, along with an overall composite estimate that includes all of a

teacher's relevant scores. The overall composite estimate for a teacher implicitly compares the teacher to all other PPS teachers whose students take at least one of the same assessments.

B. Construction of composite estimates using weights

The composite measures are obtained by combining the individual VAM estimates using a multi-step method. We first put all individual VAMs on the same scale and then averaged these components together based on the number of students contributing to each VAM and its relative precision.[25]

In the first step, we ensured that differences in the scales used by different assessments did not influence the weight that each individual VAM received in the composite. We did this by normalizing the individual VAM distributions so they all have the same SD. Prior value-added studies, including Mathematica analyses using Pittsburgh data, have found that the SD of VAM distributions can vary across measures. For example, the SD tends to be slightly larger in math than in reading. It can vary across grades within a subject, too. When not normalized, a simple average of VAM scores (for example, an average of a school's VAM scores in grade 4 and 5 based on the math PSSA) will implicitly give more weight to the distribution with the larger SD. For example, the VAM score of a top-performing school according to the measure with the larger SD will be farther from the average value, and thus larger than the VAM score of a top-performing school according to the other measure. By normalizing the VAM distributions, we avoid inadvertently putting more or less weight than desired on different VAMs.

In the second step, we produced an average that gives more weight to components that are estimated more precisely. All measures of any kind (including observation-based measures of teacher performance as well as value-added measures) are measured with some amount of uncertainty. In VAMs, the uncertainty stems both from the finite number of students included in each value-added model and from random variation in the measurement of student achievement, sometimes called statistical noise. Some students know more about an assessment's writing prompt than others; other students may be ill on the day of an exam. This variation in outcomes unrelated to student ability can lead to statistical noise in teacher VAMs unrelated to a teacher's actual value-added. Precision is reduced by statistical noise. It is increased when more students take an assessment, because there is more information available to use in measuring performance. VAM estimates tend to be less precise in subject areas where it is more difficult to measure student achievement.

Precision weighting has two characteristics that may be disadvantages. First, it gives more weight to some grades and subjects than to others. If reading scores tend to be noisier than math scores, for example, they will contribute less weight to the composite. Second, precision weighting does not capture the views of educators and policymakers about the relative importance of different outcome measures. The student assessment that produces the most precise VAM estimates may not be the one that is most important for long-term success, or the

[25] The standard errors of the composite scores were calculated using the fact that the variance of a sum is a function of the variances of each element and their covariances. We estimated the covariance of a teacher or school's VAMs using the number of students that enter multiple value-added models assigned to the same teacher or school and the covariance of the error terms ($e_{i,t,c}$ in equation (1)) across value-added models.

one that receives the most instructional time. The relative importance of different student assessments could justifiably lead PPS to choose weights that differ from the precision-maximizing weights in the future. This would produce a composite measure that has more statistical noise than a precision-weighted composite but might better reflect Pittsburgh's educational goals.

Precision weights are created based on the average precision of teacher or school VAMs for a given subject. This implies that, without further adjustments, the same weights would be used in creating the composites for teachers receiving individual VAMs for the same subjects. For example, suppose Teachers A taught fourth grade math to 60 students and fourth grade reading to 20 students, while Teacher B taught math to 20 students and reading to 60. With weights developed based only on average precision of VAMs, these teachers' math and reading VAMs would receive the same weight in their composite VAMs. This may be undesirable, as Teacher A teaches mostly reading, and Teacher B teaches mostly math.

To avoid this, our weights also took into account the number of students assigned to a teacher or school who are included in the value-added model.[26] For example, more weight was placed on Teacher A's 4th-grade PSSA reading VAM than her 4th-grade PSSA math VAM, even if the precision of the VAMs was the same. Including the number of students attributed to a teacher or school as part of the weighting in the composite ensures that the grades and subjects where teachers teach the most students (included in our models) contribute more heavily to composite estimates.

[26] For each assessment, we use the product of the number of students assigned to a teacher (school) and included in a VAM and the inverse of the average variance over all the teacher (school) value-added estimates to determine the weight that assessment receives in the teacher (school) composite estimate.

VI. SUMMARIZING PITTSBURGH SCHOOL PERFORMANCE IN THE CONTEXT OF A STATEWIDE DISTRIBUTION

PPS seeks value-added measures that allow comparisons of PPS schools to other schools in Pennsylvania. Results produced by VAMs are inherently relative to the teachers or schools included in the full data set. VAMs with statewide data (rather than only within-Pittsburgh data) therefore have the advantage that they show how Pittsburgh as a whole is performing relative to the rest of the state (in value-added terms) and how Pittsburgh's performance relative to the state changes over time. Available statewide data, however, are not as rich as Pittsburgh's own data. Most important, Pittsburgh has data on many student outcomes that are not available statewide, including CBA results, attendance, and progress in completing core courses. In addition, Pittsburgh has more data on students that can be used to improve the predictions of their likely performance. Relying exclusively on statewide data would therefore dramatically reduce the number of assessments that could be used in the VAMs, and would reduce the overall quality of the analyses.

Instead, the PPS school-level VAMs used a hybrid approach that capitalizes on the breadth of the statewide data and the richness of Pittsburgh's local data, running value-added models separately but in parallel on both data sets. To make the most of the data from district and state sources, we estimated within-district VAMs to produce fine-grained assessments of how PPS schools performed relative to each other, and we used statewide VAMs to assess where the district as a whole fell in the statewide distribution of performance. This produced a crosswalk or indirect comparison of VAM scores that allowed us to estimate the performance of each PPS school relative to the statewide average, without discarding the richer information included in the district's own data.

For teachers, in contrast, the PPS VAMs rely exclusively on the within-Pittsburgh analyses, because creating a crosswalk for teacher-level VAMs would require stronger assumptions than doing so for school-level VAMs.[27] Although all schools in Pittsburgh have data on at least one assessment that is given statewide, many teachers can be assessed only with CBAs, none of which are available statewide.

Placing Pittsburgh schools in the statewide distribution involved two steps, which will be discussed in turn. In the first step, we used data on students across Pennsylvania to estimate statewide school VAMs based on PSSA and Keystone exams. Individual VAMs were first estimated and then combined into composites. Second, we assigned a statewide NCE to each Pittsburgh school based on (1) the distribution of Pittsburgh schools in a corresponding composite statewide VAM, and (2) the finer-grained VAM rankings produced using the district-specific data. In other words, if the statewide analysis of PSSA data tells us that the top-performing Pittsburgh middle school in math had a statewide value added NCE of 90 points, then the Pittsburgh school that we identify as top-performing in Pittsburgh based on district assessments was assigned to have an NCE score of 90 points. Depending on how well it did on

[27] For example, Pennsylvania has phased out the 11th-grade PSSAs and replaced them with Keystone exams typically taken in 9th (algebra I) and 10th (literature) grades. This implies that mapping teacher performance to a statewide distribution would require the assumption that the placement of 11th- and 12th-grade Pittsburgh teachers in Pennsylvania is the same as the distribution of 9th- and 10th-grade Pittsburgh teachers in Pennsylvania.

math CBAs compared to PSSA, that school might or might not be the same one that received a VAM of 90 NCE points based only on math PSSA scores.

This process accomplishes the dual goals of PPS: that school value-added (1) be reported in the context of state performance; and (2) also incorporate assessments, like CBAs, that are offered only in Pittsburgh. Through the latter goal, we incorporated information on additional measures that are closely tied to the actual curriculum and cover a broader set of grades and subjects than could be covered by statewide assessments alone. The two-step process also allowed us to make use of finer-grained background variables available in Pittsburgh but not available statewide.

VAM measures for Pittsburgh schools included both state and locally administered assessments, but the available state distributions to which these measures can be compared were based on state assessments alone. Our method assumes that the placement of Pittsburgh schools in the statewide VAM distribution as measured by PSSA and Keystone scores is a reasonable measure of how they would place if all outcomes in the same subject were available statewide (that is, if the rest of the state had TerraNova, CBA, and PSAT results alongside PSSA and Keystone results). The rest of the chapter describes the two-step process in more depth.

A. Statewide school VAMs

Using student data from the Pennsylvania Department of Education, we estimated statewide VAMs that resemble those described in the preceding chapters as closely as possible given the available data.[28] That is, the statewide value-added models involve similar data elements and contain the same features, such as score standardization, dosage, and shrinkage. However, there are four important differences from the Pittsburgh-specific VAMs:

- **Outcomes are limited to those that are measured across Pennsylvania.** Statewide VAM analyses can include only assessments and other dependent measures for which data exist across the state (PSSAs and Keystone exams). Reliable state data do not yet exist on student attendance or core course passage, PPS's CBAs are not administered in other school districts, and the TerraNova and PSAT exams are given only to a subset of students outside PPS.
- **Sample includes more students.** The sample size for each statewide VAM was substantially larger than it was when estimating a VAM based on Pittsburgh students only. The eligible sample for each individual statewide VAM included all students with data on a particular outcome measure. For example, a statewide VAM could include all Pennsylvania students with a score on the grade 6 math PSSA. This larger sample size led to more precise value-added estimates because we were able to measure the relationships between student characteristics and achievement more precisely.
- **Differences in student-level control variables.** The state data contain much of the same student information that we use in the PPS VAMs, though the alignment is not perfect. Specifically, in the statewide analyses we could not include information on gifted

[28] Assessment data come from the Bureau of Assessment and Accountability. All other student data come from the Pennsylvania Information Management System.

participation, course type, prior-year absences, prior-year suspensions, or prior full-year district membership.[29] To limit the potential bias associated with including fewer background characteristics, we added a control for students' own test scores in the same subject from the second prior grade (as a third baseline score).

- **Less exact dosage measure.** Because data on mid-year student transfers are not currently available at the statewide level, school dosage measures are less exact in the statewide VAMs than in the PPS VAMs. We determined the number of schools a student attended during the year and assume an equal dosage between them.

B. Assigning a state Value-Added NCE to results based on Pittsburgh data

Based on how the distribution of performance in Pittsburgh falls relative to the state, our final step was to assign a state value-added NCE to the Pittsburgh VAM estimates. The distribution of PPS-specific VAM estimates was adjusted to match the distribution of estimated PPS value added in the statewide analyses. This process attempts to make the most of the available information: statewide VAM estimates are used to determine the general ranking of PPS schools' performance in the state, and PPS-specific VAM estimates use finer-grained data—including more student-level variables and additional outcomes—to provide a better indication of where each PPS school falls in the district-wide distribution.[30]

All schools, regardless of grade configuration, were placed in the same distribution when determining the statewide NCE rank. This means that a school's NCE rank is relative to all schools in Pennsylvania. Since each VAM was estimated separately by assessment and grade and as a result of the normalization of VAM estimates described in Section V.B, a school's place in the statewide distribution was almost entirely determined by its performance relative to other schools that serve the same grades and administer the same assessments. The multiple possible overlapping grade ranges of Pittsburgh schools (K-5, K-8, 6-8, 6-12, and 9-12) made it difficult to compare schools only to other schools with the same grade configurations when determining the NCE rank. We therefore placed schools into one statewide distribution to ensure that all schools with overlapping grade ranges are compared to each other.

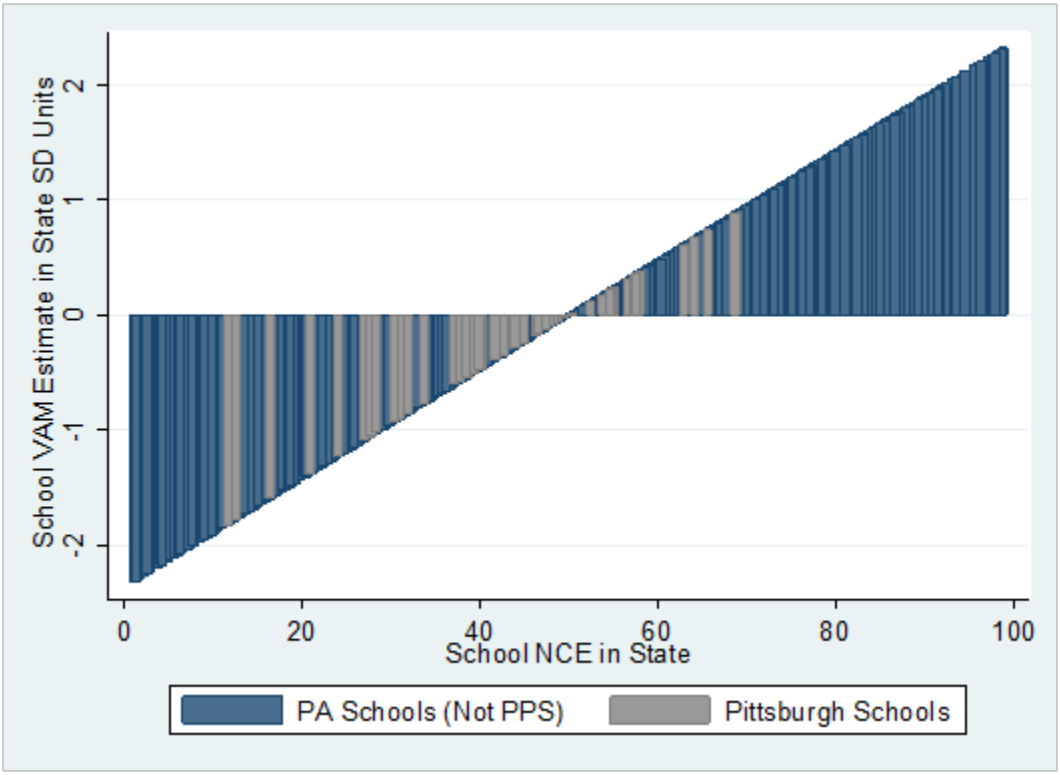
Figure VI.1 illustrates where PPS schools fall in the statewide distribution on the overall composite value-added measure. More Pittsburgh schools rank below the statewide average in terms of overall value-added, though a few exceed the average. The average Pittsburgh school received an NCE of 42 (roughly the 35th percentile) in the statewide distribution using data from 2011 to 2013. The composite value-added scores of Pittsburgh schools range from 11 to 69

[29] We excluded the gifted program participation field in PIMS from the state VAM analyses because we are concerned about its validity. The data suggest that no Pittsburgh students participate in the gifted program. In contrast, Pittsburgh's own data indicate that Pittsburgh students participated in the gifted program at a rate double the statewide average. We opted against replacing PIMS gifted data for PPS with RTI information, because we were concerned with the validity of gifted information in other districts as well.

[30] Although school VAMs are adjusted using the statewide data, we used only district data when determining the width of the confidence intervals around these VAMs. This accurately reflects the lower precision we have in the smaller, district-wide data set. Confidence intervals calculated using data from across the state were too narrow and overstated our true level of precision.

normal curve equivalent points (or the 5th to 82nd percentiles), spanning much of the statewide distribution.

Figure VI.1. Distribution of composite school VAM estimates in Pennsylvania, 2011–13



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

Note: This figure displays the NCE rank of Pittsburgh schools relative to other schools in Pennsylvania based on their composite VAM. VAM scores are top and bottom coded such that they range between 1 and 99 NCEs.

As noted earlier, fewer outcome measures are available at the state level than are available in Pittsburgh. Table VI.1 lists the assessments available statewide, categorized by subject-level composite. PPS-specific composites that include local assessments alongside PSSAs and Keystones were mapped to statewide composites as indicated in Table VI.1.

Comparing this table to Table V.1 reveals how we map district composites to state composites containing different test scores. For example, the PPS elementary math composite included VAMs based on Grade 2 TerraNova math and Grades 3, 4, and 5 PSSA math scores. This was matched to a statewide math composite that included only PSSA scores for grades 4 and 5. In middle school grades, the state composites included only PSSA VAMs, whereas the district composites included both PSSA and CBA VAMs. Similarly, district-level composites VAMs at the high school level incorporated information from PSSAs, Keystone exams, CBAs, and the PSAT; state-level composites used only the first two of these exams. Finally, note that no social studies exams are available statewide. Thus, the district-level VAM composites in social

studies were mapped to state-level composites created using reading, writing, and literature assessments available across the state.

Table VI.1. The composition of statewide composites, 2011–13

	Elementary school grades K to 5	Middle school grades 6 to 8	High school grades 9 to 12
Test-based measures			
Math composite	PSSA math (4,5)	PSSA math (6,7,8)	PSSA math (11) Keystone algebra I (9)
English/language arts composite	PSSA reading (4,5) PSSA writing (5)	PSSA reading (6,7,8) PSSA writing (8)	PSSA reading (11) PSSA writing (11) Keystone literature (10)
Science composite	PSSA science (4)	PSSA science (8)	PSSA math (11) Keystone algebra I (9)
Social studies composite	n/a	PSSA reading (6,7,8) PSSA writing (8)	PSSA reading (11) PSSA writing (11) Keystone literature (10)

Note: The grade level of the majority of students follows each assessment in parentheses. Eleventh-grade PSSAs were used to assess high schools across the state in 2011–12, and Keystone algebra I and literature were used during 2012–13. The two-year school VAMs for 2011–2013 used a composite of the Keystone and PSSA scores. In subsequent years, only the Keystone assessments will be used. The grade 8 PSSA science VAM incorporates data form 2011-12 only.

Using these mappings, we found that Pittsburgh schools tended to perform worse overall than the state average. Altogether, 30 percent of Pittsburgh schools had overall composite effectiveness scores statistically different from the average school statewide. These schools included one with above-average and 14 with below-average statewide performance.

C. Other school performance metrics: PVAAS and SPP

In addition to VAMs, two other measures are currently being used for school evaluation in Pittsburgh. These measures—provided by the Pennsylvania Department of Education—are the Pennsylvania Value Added Assessment System (PVAAS) and the School Performance Profile (SPP). In this section, we discuss the similarities and differences between school VAMs and these two school performance measures.

PVAAS is similar to school VAMs in that it provides PPS schools with estimates of how their value added compares to that of other schools in the state. There are four key differences between the PVAAS model and the Pittsburgh value-added model, however.[31]

[31] See Wright et al. (2010) for a technical report describing the PVAAS model. There are also a number of differences in the methods used to estimate the Pittsburgh and PVAAS models. In particular, the Pittsburgh model incorporates an explicit measurement error adjustment for prior test scores, whereas PVAAS relies on the inclusion of multiple years of prior scores to account for measurement error. The Pittsburgh model also treats the school effects as fixed, whereas the PVAAS model treats them as random.

1. PVAAS measures growth in student achievement using only assessments that are available statewide (PSSA and Keystone exams). Pittsburgh VAMs include these state assessments, but they also use a broader array of assessments administered in PPS (TerraNova, CBA, and PSAT exams).
2. Pittsburgh VAMs average over school performance during the past two academic years, whereas PVAAS measures school performance only during the past year.
3. PVAAS includes controls for multiple years of prior student test scores in its models, whereas the Pittsburgh VAMs include test scores in multiple subjects from the prior year only (except at the high school level).
4. Pittsburgh VAMs account for student level demographic characteristics and class averages of these variables, whereas PVAAS only includes prior test scores as control variables.

Despite these differences in methodology, fundamentally PVAAS and Pittsburg school VAMs are measuring school contributions to student achievement accounting for prior test scores. Therefore, these two measures tend to produce results that are related to each other. In a previous report, we compared PVAAS estimates with one-year estimates from Pittsburgh school VAMs and found results to be positively correlated. We found an average correlation of about 0.70 across math, reading, science, and writing VAMs (see Appendix A.2 of Johnson et al. 2012 for details). Although there is a positive correlation between the Pittsburgh and the PVAAS numeric VAM scores, the PVAAS results that receive the most attention—the color codes that measure performance against an academic growth standard set by Pennsylvania—tend to differ substantially from Pittsburgh school VAM estimates. This is in part because PVAAS bases the color coding on an absolute-growth standard, whereas school VAMs compare Pittsburgh schools to statewide average performance.

The SPP system provides a metric that includes information on student test score growth as well as other measures of achievement. SPP is related to PVAAS in that 40 percent of the SPP score is based on a school's PVAAS score. The other 60 percent is made up of measures based on student achievement levels. These measures include the percentage of students who earned proficient or advanced on state assessments, participation and performance on the SAT and ACT exams, graduation rate, and attendance rate. Therefore, SPP can be thought of as a measure that captures primarily the achievement levels of students, whereas PVAAS and Pittsburgh school VAMs are based entirely on student achievement growth.

VII. THE DISTRIBUTION OF TEACHER AND SCHOOL VALUE ADDED IN PPS

In this chapter, we examine the distribution of subject-level and composite school and teacher VAMs across PPS. This analysis allows us to better understand how the efficacy of teachers and schools varies across the district. It also allows us to quantify differences in VAMs, relating these gaps to differences in learning. By examining the distribution of VAMs, we can explore the extent to which school and teacher assignment can influence student achievement.

A. Teacher VAM results

The summary results for the teacher VAMs are displayed by grade and assessment in Table VII.1. On average across all assessments, using a 95 percent confidence interval, we can distinguish 40 percent of teachers from the PPS average. The dispersion of value-added estimates, measured by the SD of the VAMs, varies by grade, subject, and assessment.[32] At the extremes, the 90th-percentile teacher raised achievement on the 8th-grade physics CBA by 0.58 SDs compared to the average teacher, while the 90th-percentile teacher raised achievement on the 11th-grade reading PSSA by 0.10 SDs.

A difference of a given size should be interpreted within the scope of the progress students typically make in a subject over the year. That is, suppose a student typically scores 1.00 SDs higher on a test when it is administered at the end of the year compared to the beginning. Then a 0.25-SD difference in test scores amounts to 25 percent of what a student typically learns in a year. This has a meaning very different from an assessment where a typical student's scores improve by only 0.50 SDs over the course of a year. In this case, a 0.25-SD difference in test scores amounts to 50 percent of what students are expected to learn during the year. Thus, converting SD differences to differences in years of learning can help us better understand the meaning behind differences in VAMs.

Hill et al. (2008) provide estimates of expected test score gains by grade that we used to convert SD differences to differences in years of learning. These estimates, based on results from seven nationally normed vertically scaled assessments, provide valuable information but should be interpreted with caution. The accuracy of our conversion to years of learning growth depends on an assumption that the variance of student achievement on TerraNova and PSSA exams in Pittsburgh is approximately equivalent to the variance of student achievement on the assessments analyzed by Hill and colleagues. Under this assumption, Figure VII.1 shows how the teacher value-added estimates for TerraNova and PSSA outcomes can be described in terms of the proportion of the average amount of learning typically achieved by a student in that grade and subject. CBA outcomes were not included in this table because they are specific to Pittsburgh and are likely not comparable to the nationally normed assessments used to estimate gains of typical students.

On average across grades and subjects, a typical student with a 90th-percentile teacher learns approximately an additional 51 percent of a typical year of learning growth, relative to

[32] Estimates from value-added models are likely to overstate the SD of true teacher effectiveness. To correct for this, we report the sampling-error-adjusted SD, as advocated by Aaronson et al. (2007) and detailed in Morris (1983).

how much is learned by a student with the median Pittsburgh teacher. This number is similar to the average reported in Johnson et al. (2012): 57 percent. The difference between the teachers at the 90th and 50th percentiles varies by grade and subject, in part because students make larger gains in some grade levels and subjects than in others.[33] Hill et al. (2008) found that annual gains are largest at lower grade levels, with average gains between 1st and 2nd grade of 0.97 SDs in reading and 1.03 SDs in math. In Pittsburgh, the 90th-percentile teacher increased 2nd-grade math achievement by 0.32 SDs more than the average teacher, or about 31 percent of what a typical 2nd-grade student would be expected to gain in math during the school year. Meanwhile, the 90th-percentile teacher increased 6th-grade achievement by 0.22 SDs on the reading PSSA compared to the average Pittsburgh teacher. This equates to about 69 percent of what a typical 6th-grade student would be expected to gain in reading during the school year.

Table VII.1. Teacher VAM results, by outcome 2010–13

Outcome	Grade	Adj. R-squared	Years of teaching	Teachers	Difference between 90th- and 50th-percentile in z-score units	SD of teacher effects	Mean standard error	Share of VAMs statistically significant (95% CI)
TerraNova Math	2	0.76	1	60	0.32	0.25	0.10	0.32
TerraNova Reading	2	0.70	1	63	0.24	0.19	0.10	0.17
PSSA Math	3	0.74	2	76	0.27	0.21	0.09	0.34
PSSA Reading	3	0.73	2	81	0.26	0.20	0.10	0.17
PSSA Math	4	0.78	3	90	0.22	0.17	0.08	0.31
PSSA Reading	4	0.74	3	88	0.22	0.17	0.08	0.31
PSSA Science	4	0.76	3	71	0.24	0.19	0.08	0.46
PSSA Math	5	0.84	3	76	0.23	0.18	0.07	0.41
PSSA Reading	5	0.74	3	88	0.21	0.16	0.08	0.33
PSSA Writing	5	0.54	3	89	0.35	0.27	0.11	0.44
CBA Earth Science	6	0.61	3	53	0.49	0.38	0.09	0.62
CBA Math	6	0.63	3	59	0.28	0.22	0.09	0.39
CBA Reading	6	0.59	3	64	0.25	0.19	0.09	0.33
CBA Social Studies	6	0.63	1	18	0.42	0.33	0.10	0.50
PSSA Math	6	0.80	3	64	0.20	0.15	0.07	0.39
PSSA Reading	6	0.73	3	93	0.22	0.17	0.09	0.23
CBA Life Science	7	0.72	2	34	0.44	0.34	0.09	0.53
CBA Math	7	0.60	3	63	0.31	0.24	0.09	0.41
CBA Physical Science	7	0.65	1	19	0.42	0.32	0.08	0.58
CBA Reading	7	0.59	3	65	0.21	0.17	0.09	0.23
CBA Social Studies	7	0.58	1	14	0.48	0.37	0.09	0.57

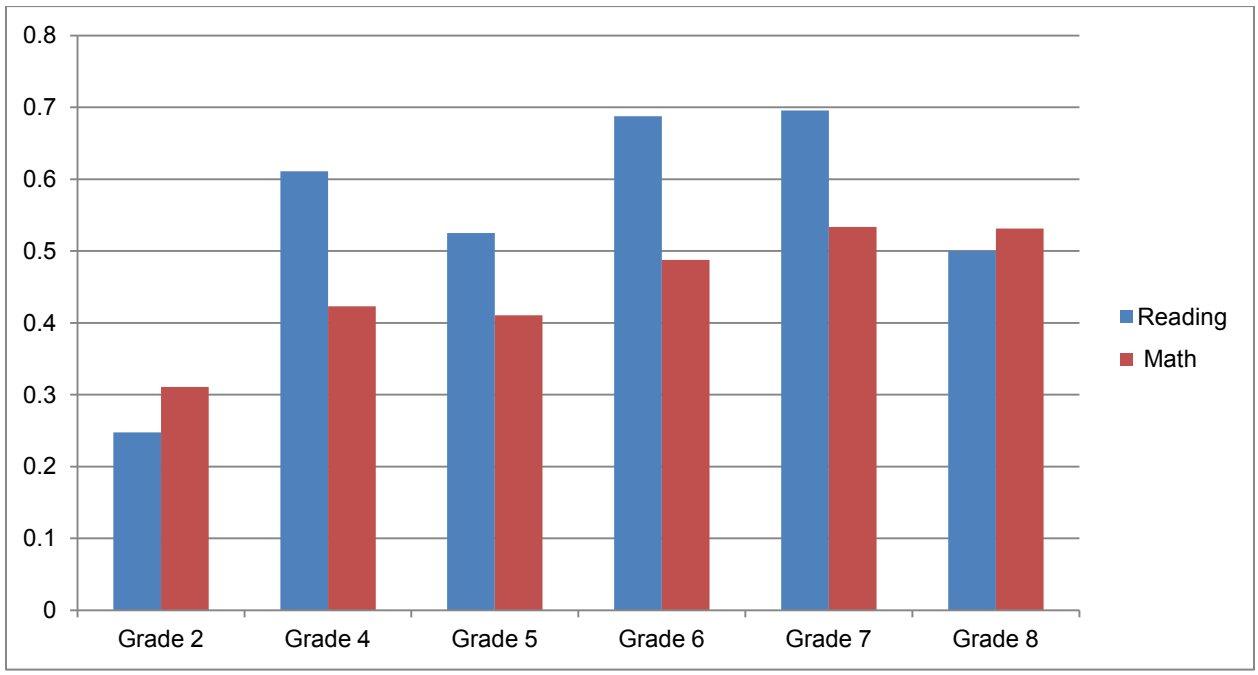
[33] To make growth comparisons that are as similar as possible to those in Hill et al. (2008), we translate measures into years of learning gains only for VAMs that use the same assessment as a baseline and outcome variable.

Table VII.1 *(continued)*

Outcome	Grade	Adj. R-squared	Years of teaching	Teachers	Difference between 90th- and 50th-percentile in z-score units	SD of teacher effects	Mean standard error	Share of VAMs statistically significant (95% CI)
PSSA Math	7	0.82	3	67	0.16	0.12	0.06	0.25
PSSA Reading	7	0.76	3	88	0.16	0.12	0.07	0.20
CBA Math	8	0.41	3	48	0.40	0.31	0.12	0.46
CBA Physical Science	8	0.63	1	18	0.34	0.27	0.08	0.61
CBA Physics	8	0.65	2	30	0.58	0.46	0.09	0.80
CBA Reading	8	0.60	3	57	0.25	0.20	0.09	0.40
CBA US History	8	0.64	3	38	0.43	0.34	0.09	0.53
PSSA Math	8	0.82	3	64	0.17	0.13	0.07	0.27
PSSA Reading	8	0.76	3	66	0.13	0.11	0.07	0.08
PSSA Science	8	0.74	2	32	0.21	0.16	0.07	0.28
PSSA Writing	8	0.55	3	62	0.32	0.25	0.10	0.40
CBA Algebra I	9	0.43	3	44	0.34	0.27	0.11	0.30
CBA Biology	9	0.49	3	24	0.37	0.29	0.09	0.58
CBA Civics	9	0.66	3	31	0.50	0.39	0.08	0.84
CBA ELA 1	9	0.50	3	47	0.32	0.25	0.10	0.30
Keystone Algebra I	9	0.56	1	16	0.19	0.15	0.08	0.25
Spanish Multimode Level 1	9	0.41	3	5	0.19	0.15	0.17	0.00
CBA Chemistry	10	0.48	3	25	0.36	0.28	0.11	0.48
CBA Civics	10	0.60	3	34	0.46	0.36	0.11	0.65
CBA ELA II	10	0.53	3	44	0.20	0.16	0.09	0.14
CBA Geometry	10	0.51	3	35	0.31	0.24	0.10	0.31
French Multimode Level 2	10	0.56	3	11	0.35	0.27	0.20	0.09
Keystone Literature	10	0.62	1	19	0.19	0.15	0.09	0.21
Spanish Multimode Level 2	10	0.43	3	19	0.39	0.30	0.12	0.74
CBA Algebra II	11	0.46	3	49	0.44	0.34	0.12	0.39
CBA ELA III	11	0.48	3	39	0.25	0.20	0.11	0.21
CBA Physics	11	0.57	3	27	0.48	0.38	0.13	0.74
CBA US History	11	0.50	3	25	0.50	0.39	0.13	0.52
PSSA Reading	11	0.64	2	34	0.10	0.08	0.07	0.06
PSSA Writing	11	0.49	2	34	0.13	0.10	0.10	0.06
CBA ELA IV/AA Lit	12	0.53	3	39	0.18	0.14	0.10	0.23

Source: Authors' calculations based on data provided by PPS.

Figure VII.1. Teacher VAM results for nationally normed tests expressed in fractions of a year of learning: Difference between median teacher and 90th-percentile teacher in Pittsburgh



Source: Authors' calculations based on data provided by PPS.

Notes: Grade 2 assessments are TerraNova tests and assessment for grades 4-8 are PSSA tests. The difference in effectiveness in standard deviation units is converted to years of learning using the numbers reported in Hill et al. (2008). Other grades are omitted from this figure, as they use different baseline and outcome assessments (for example, grade 3 PSSA VAMs are estimated using baseline grade 2 TerraNova scores).

The overall composite effectiveness score for teachers was calculated using the test-based VAMs described by Table VII.1. Overall composite effectiveness measures are available for 775 teachers, of whom 37 percent can be distinguished statistically from typical performance using a 95 percent confidence level.

B. School VAM results

In this section, we first discuss the results of the district-wide school VAM estimates based on individual assessments and grouped by grade level. Although actual composite value-added estimates were based on statewide VAMs, the district analysis is important because it determines the rank of schools within the district. We start by presenting summary results for the wide range of assessments that are used in district VAMs. We then turn our analysis to composite measures and explore the extent to which PPS schools can be distinguished from the average school in Pennsylvania.

1. School VAM results for grades 2-5

The results of the district-wide school level VAMs for grades 2 to 5 are presented in Table VII.2. The models for 3rd to 5th grade were based on two years of data; those for 2nd grade were based on a single year of data. Across all assessments in this grade range, the 90th-percentile

school raises achievement by between 0.14 and 0.34 SDs compared to the average school. These findings are similar to those presented in Johnson et al. (2012). For the 2010–11 school year, that report listed differences ranging from 0.14 to 0.33 SDs.

Differences can be interpreted in terms of the expected gains of a typical student in each grade and subject, based on the estimates reported by Hill et al. (2008). For example, a 4th-grade student at the 90th-percentile school learns, on average, about 33 percent more in a year in math than a typical year's worth of learning. A test of the joint significance of results for each VAM confirms that results are not merely random distributions.

Table VII.2. School VAM results for grades 2 to 5, by outcome

					Difference between 90th- and 50th-percentile schools			
Outcome	Grade	Adj. R-squared	Students	Schools	In z-score units	In fractions of a year of learning	Mean standard error	Share of VAMs statistically significant
TerraNova Math	2	0.75	1498	34	0.34	0.33	0.11	0.44
TerraNova Reading	2	0.70	1494	34	0.25	0.26	0.11	0.15
PSSA Math	3	0.74	3000	34	0.23	NA	0.07	0.50
PSSA Reading	3	0.72	2963	34	0.18	NA	0.07	0.15
PSSA Math	4	0.81	3032	34	0.17	0.33	0.06	0.35
PSSA Reading	4	0.77	3018	34	0.14	0.38	0.06	0.18
PSSA Science	4	0.77	2995	34	0.20	NA	0.07	0.38
PSSA Math	5	0.84	3090	34	0.17	0.30	0.06	0.38
PSSA Reading	5	0.75	3083	34	0.23	0.58	0.07	0.56
PSSA Writing	5	0.52	3039	34	0.22	NA	0.08	0.29

Source: Authors' calculations based on data provided by PPS.

Note: Difference between 90th- and 50th-percentile school in terms of one year of learning is based on estimates from Hill et al. (2008).

2. School VAM results for grades 6-8

The results of the school-level VAMs for grades 6 through 8 are presented in Table VII.3. Most of these models were based on two years of data, though many of the social studies and science outcomes used a single year of data. A student at the 90th-percentile PPS school learns, on average, between 41 and 57 percent more than the amount learned in a typical PPS school in the middle school subjects. These estimates are again similar to those reported for the 2010–11 school year in Johnson et al. (2012), in which students in the 90th-percentile PPS school learned 35 to 60 percent more than those at the average PPS school.

There is more variation in school effects across assessments in middle schools than in elementary schools. For example, the 90th-percentile school raises achievement on the 6th-grade social studies CBA by 0.62 SDs compared to the district average, while the 90th-percentile school on the 7th-grade reading PSSA improves results by only 0.10 SDs relative to a typical PPS school. In the elementary grades, the range of gains was smaller (0.14 to 0.34 SDs).

3. School VAM results for grades 9-12

The results of the school-level VAMs for grades 9 through 12 are presented in Table VII.4. The models examining CBAs used two years of data, while those for PSATs, PSSA, and Keystone exams used a single year of data.[34] We do not include a conversion to years of learning growth for high school VAMs, because baseline scores generally do not come from the same assessment.

As in lower grades, the range of school effects varies across subjects. The 90th-percentile PPS school raises achievement on the biology CBA by 0.57 SDs compared to the average PPS school, but this difference is only 0.07 SDs on the 11th-grade reading PSSA. The SDs are similar to those reported in Johnson et al. (2012). In the earlier report, differences between the 90th percentile and average PPS schools ranged from 0.03 to 0.45 SDs.

[34] Because of the timing of the exams, PSAT value-added models included only one year of data. PSATs are taken early in the fall of each school year, so student performance on the PSAT is attributed to the school during the previous academic year. For example, PSATs taken in the fall of 2013 are attributed to the school a student was enrolled in during the 2012–13 academic year. However, these scores were not available in time for inclusion in the model estimation, so only one year of PSAT data (exams taken in the fall of 2012) were available to use in VAMs.

Table VII.3. School VAM results for grades 6 to 8, by outcome

Outcome	Grade	Adj. R-squared	Students	Schools	Years of data	Difference between 90th- and 50th-percentile schools		Mean standard error	Share of VAMs statistically significant
						In z-score units	In fractions of a year of learning		
CBA Earth Science	6	0.59	2586	25	2	0.49	NA	0.09	0.68
CBA Math	6	0.63	2562	24	2	0.24	NA	0.08	0.33
CBA Reading	6	0.61	2190	25	2	0.20	NA	0.08	0.24
CBA Social Studies	6	0.66	711	16	1	0.62	NA	0.22	0.38
PSSA Math	6	0.80	2827	25	2	0.20	0.48	0.06	0.52
PSSA Reading	6	0.73	2869	25	2	0.16	0.50	0.06	0.40
CBA Life Science	7	0.72	1169	21	1	0.50	NA	0.11	0.62
CBA Math	7	0.62	2465	23	2	0.24	NA	0.08	0.39
CBA Physical Science	7	0.67	1131	19	1	0.39	NA	0.14	0.37
CBA Reading	7	0.59	2395	24	2	0.08	NA	0.06	0.00
CBA Social Studies	7	0.60	766	15	1	0.66	NA	0.29	0.07
PSSA Math	7	0.81	2840	24	2	0.16	0.53	0.05	0.46
PSSA Reading	7	0.76	2975	24	2	0.10	0.43	0.05	0.42
CBA Math	8	0.42	1605	23	2	0.45	NA	0.13	0.48
CBA Physical Science	8	0.64	1093	19	1	0.34	NA	0.13	0.32
CBA Physics	8	0.70	1151	22	1	0.56	NA	0.12	0.55
CBA Reading	8	0.61	2433	24	2	0.15	NA	0.07	0.13
CBA US History	8	0.68	2358	24	2	0.25	NA	0.08	0.33
PSSA Math	8	0.78	2149	23	2	0.18	0.57	0.06	0.39
PSSA Reading	8	0.77	2859	24	2	0.11	0.41	0.06	0.13
PSSA Science	8	0.79	1368	23	1	0.15	NA	0.08	0.17
PSSA Writing	8	0.55	2798	24	2	0.32	NA	0.08	0.46

Source: Authors' calculations based on data provided by PPS.

Note: Difference between 90th- and 50th-percentile school in terms of one year of learning is based on estimates from Hill et al. (2008), as described in the text above.

Table VII.4. School VAM results for grades 9 to 12, by outcome

Outcome	Grade	Adj. R-squared	Students	Schools	Years of data	Difference between 90th- and 50th-percentile schools in z-score units	Mean standard error	Share of VAMs statistically significant
CBA Algebra I	9	0.40	1315	9	2	0.40	0.09	0.56
CBA Biology	9	0.47	1374	7	2	0.59	0.09	0.71
CBA Civics	9	0.64	1491	9	2	0.54	0.08	0.78
CBA ELA 1	9	0.47	1622	9	2	0.24	0.08	0.44
Keystone Algebra I	9	0.57	701	9	1	0.27	0.12	0.11
CBA Chemistry	10	0.51	970	7	2	0.24	0.10	0.71
CBA Civics	10	0.57	1267	8	2	0.27	0.09	0.50
CBA ELA II	10	0.50	1521	9	2	0.17	0.07	0.44
CBA Geometry	10	0.45	1357	9	2	0.31	0.08	0.56
Keystone Literature	10	0.62	839	8	1	0.15	0.09	0.00
PSAT Math	10	0.58	922	9	1	0.16	0.05	0.33
PSAT Reading	10	0.51	812	9	1	0.15	0.06	0.44
PSAT Writing	10	0.47	812	9	1	0.20	0.06	0.56
CBA Algebra II	11	0.41	1441	9	2	0.34	0.10	0.44
CBA ELA III	11	0.41	1053	8	2	0.26	0.10	0.25
CBA Physics	11	0.55	1007	8	2	0.51	0.10	0.63
CBA US History	11	0.51	1027	8	2	0.57	0.11	0.63
PSAT Math	11	0.64	722	9	1	0.24	0.06	0.56
PSAT Reading	11	0.59	703	9	1	0.26	0.07	0.67
PSAT Writing	11	0.57	703	9	1	0.27	0.07	0.89
PSSA Math	11	0.56	451	7	1	0.40	0.10	0.57
PSSA Reading	11	0.58	528	7	1	0.07	0.08	0.00
PSSA Science	11	0.60	515	7	1	0.12	0.11	0.14
PSSA Writing	11	0.49	515	7	1	0.35	0.13	0.29
CBA ELA IV/AA Lit	12	0.54	1223	8	2	0.17	0.08	0.25

Source: Authors' calculations based on data provided by PPS.

4. Composite school VAM results

After estimating all the assessment-level school VAMs, we combined the test-based outcomes into four subject-level composites, an overall test-based composite, and two non-test composites (see Chapter V). We then mapped the within-district performance of Pittsburgh schools to their performance relative to other schools in the state (see Chapter VI). The number of PPS schools in each grade range receiving composite VAMs, as well as the fraction of PPS schools distinguishable from the state average, are reported in Table VII.5. We report these statistics for the attendance and core course pass rate VAMs as well, though for the non-test VAMs, comparisons were made to the average PPS school, and not the average school in

Pennsylvania, because of the limited availability of data on non-test outcomes outside the district.

Overall, 46 percent of PPS schools could be statistically distinguished from the average Pennsylvania school based on math scores, 20 percent based on reading scores, 44 percent based on science scores, and 37 percent based on social studies scores. Using the test-based composites, 30 percent of PPS schools are significantly better or worse than the average school in the state. Based on attendance and core course pass rates, we can distinguish 48 percent and 44 percent (respectively) of PPS schools from the average school in Pittsburgh.

Several patterns emerge for the composite estimates. We see that fewer of the PPS elementary schools can be distinguished from the state average school compared to schools incorporating higher grades. For example, 14 percent of PPS K-5 can be distinguished from the state average on the test-based composite compared to 29 percent or higher for schools serving other grade ranges. There is also variation by subject; we were more likely to be able to distinguish a school from average based on the math composite than based on the reading composite.

Table VII.5. Test-based composite school VAM results

		Statistically significant effects (95% CI)						
School type	Schools	Math	Reading	Science	Social studies	Overall test-based composite	Attendance rate	Core-course pass rate
K-5	22	0.32	0.05	0.14	NA	0.14	0.36	NA
K-8	12	0.67	0.33	0.67	0.27	0.42	0.27	NA
6-8	7	0.43	0.29	0.86	0.29	0.29	0.86	NA
6-12	5	0.40	0.20	0.60	0.60	0.60	0.60	0.40
9-12	4	0.75	0.50	0.50	0.50	0.50	1.00	0.50

Source: Authors' calculations based on data provided by PPS.

This page has been left blank for double-sided copying.

VIII. THE RELATIONSHIP BETWEEN STUDENT AND CLASSROOM CHARACTERISTICS AND TEST SCORES

PPS value-added models account for a number of student and classroom characteristics to avoid unfairly penalizing or rewarding teachers or schools based on the underlying ability or demographic composition of their students. Although the chief purpose of including student and classroom characteristics is to promote the fairness and validity of the VAMs, the relationships between these factors and test scores may be of independent interest. Examining these relationships can help us to understand how and why achievement varies across PPS students. This understanding can further inform decisions made by PFT, PPS, and other key stakeholders.

The estimated relationships take the form of regression coefficients. For binary characteristics, the coefficient tells us the difference in predicted test scores between students with a given characteristic (for example, gifted students) and those without that characteristic (non-gifted students), holding all else equal. For example, on average across the value-added models, we find that gifted students typically score 0.08 SDs higher than non-gifted students, holding all other student and classroom characteristics equal. This means that when comparing two students in the same classroom who are otherwise identical (based on their characteristics), the gifted student is predicted to score 0.08 standard deviations higher on the post-test. For characteristics that have multiple categories (such as race), these coefficients tell us the difference in predicted test scores between students in a given category (for example, white students) and those in the reference group (African American students), holding all else equal. Finally, regression coefficients for continuous numerical characteristics (for example, share of students in the classroom that are white) tell us the change in test scores associated with a one-unit change in the variable, holding all else equal.

Table VIII.1 describes the regression coefficients associated with the different student characteristics included in our teacher value-added models.[35] We summarize the average relationship between the control variable and test scores across all the value-added models. The first column contains the average regression coefficient. The second and third columns list the proportion of value-added models in which a characteristic is associated with statically significantly higher or lower test scores. All estimates in this table are based on models that account for other student characteristics, classroom characteristics, and prior test scores (see Chapter III for details).

The estimates indicate relationships between student characteristics and outcomes that are consistent with the literature. After accounting for prior test scores and other characteristics, white and Asian students often have significantly higher achievement than African American students (the reference group). Male students perform slightly worse than females on average; however, in some subjects they tend to do significantly better. Students from low-income families, as measured by free- or reduced-price-lunch status, also perform significantly worse than other students in 35 percent of the value-added models we estimated.

[35] School value-added models yield largely similar estimates of the coefficients on control variables.

Interestingly, English-language learners typically exhibit above-expected test scores. Although this may seem counterintuitive, it need not be, because our models account for prior-year test scores. Thus, English-language learners may have higher achievement than otherwise equivalent students as they learn English and catch up to their peers. Likewise, it appears that grade repeaters have higher test scores than other students. Students who missed many days in the past school year and those who were suspended in the past school year tend to have lower achievement than other students. In addition to accounting for the relationship between prior-year attendance and current year scores, these characteristics may also capture an individual's tendency to miss school or be in trouble. That is, students who were absent or suspended for many days in the past school year are more likely to miss school in the current year, potentially bringing down current-year test scores.

Table VIII.1. Relationship between student characteristics and test scores: Evidence from teacher Value-Added Models

Control variable	Coefficient in student-level z-score units		
	Average across Value-Added Models	Share of Value-Added Models where significant and positive (95% level)	Share of Value-Added Models where significant and negative (95% level)
Race (African American Is reference category)			
White	0.06	0.48	0.02
Hispanic	0.07	0.15	0.00
Asian	0.16	0.44	0.00
Other race	0.03	0.15	0.04
Ever applied for special services	-0.01	0.02	0.02
Ever applied to magnet school	0.02	0.25	0.02
Male	-0.02	0.29	0.35
Lunch program	-0.04	0.00	0.35
English-language learner	0.15	0.23	0.04
Gifted	0.08	0.48	0.02
Moved schools last year	-0.02	0.04	0.02
Past year proportion absent	-0.27	0.02	0.27
Past year proportion suspended	-0.80	0.00	0.15
In PPS all of past year	-0.03	0.06	0.18
Age in years	-0.04	0.00	0.31
Behind age-appropriate grade level	0.00	0.00	0.02
Specific learning disability	-0.07	0.04	0.27
Speech or language impairment	0.00	0.04	0.04
Emotional disturbance	-0.05	0.02	0.19
Intellectual disability	-0.02	0.04	0.16
Autism	0.07	0.17	0.06
Physical/sensory impairment	0.06	0.04	0.04
Other impairment	-0.11	0.04	0.21
Repeating grade	0.06	0.24	0.06
Pittsburgh Scholars Program	0.05	0.33	0.14
Advanced Placement Program	0.05	0.20	0.20
Center for Advanced Studies Student	-0.03	0.00	0.00

Source: Authors' calculations based on data provided by PPS.

Note: Estimated relationships from all teacher-level, test-based, value-added models including the relevant characteristic.

In addition to student-level characteristics, we included classroom-average characteristics in our value-added models. Table VIII.2 summarizes these estimated relationships. Far fewer of these coefficients are consistently significant and of the same sign. For example, a 10-percentage-point increase in the share of students that are gifted in a class tends to increase test scores by about 1 percent of one SD. The relationship is significant and positive in 17 percent of value-added models but significant and negative in 7 percent. Overall, student-level characteristics tend to have more consistent relationships with predicted test scores than classroom average characteristics.

Table VIII.2. Relationship between classroom characteristics and test scores: Evidence from teacher Value-Added Models

	Coefficient in student-level z-score units		
Control variable	Average across Value-Added Models	Share of Value-Added Models where significant and positive (95% level)	Share of Value-Added Models where significant and negative (95% level)
Share white	0.02	0.22	0.10
Share Hispanic	0.21	0.15	0.02
Share Asian	0.00	0.07	0.07
Share other race	0.01	0.12	0.02
Share male	-0.01	0.12	0.17
Share in lunch program	0.00	0.10	0.15
Share English language learners	-0.06	0.05	0.10
Share gifted	0.12	0.17	0.07
Share with any disability	-0.05	0.12	0.15
Average past year proportion absent	0.19	0.24	0.17
Average past year proportion suspended	-1.06	0.10	0.15
Share in PPS all of past year	0.05	0.20	0.07
Class Size	0.00	0.05	0.24
Past year math score	-0.03	0.05	0.17
Past year reading score	0.06	0.17	0.02

Source: Authors' calculations based on data provided by PPS.

Notes: Estimated relationships from all teacher-level, test-based value-added models including the relevant characteristic. Past year scores are TerraNova assessments from grade 2 for grade 3 assessments, PSSA scores in the previous year for assessments from grades 3-8, and 8th-grade PSSA for high school assessments.

IX. APPLICATIONS TO REWARDS AND RECOGNITION OPPORTUNITIES

In collaboration with the PFT, PPS has developed programs to recognize and reward the schools, teams, and individuals that are producing large improvements in outcomes for their students. Two programs—both developed by collaborative groups of principals, teachers, district staff, and PFT staff based on plans described in the 2010 collective bargaining agreement—use value-added composite measures in calculating those awards. The first is a team-based award for Promise-Readiness Corps teams in the high schools; the second is a school-based award under the STAR program. Details about the Promise-Readiness Corps value-added model are in a separate report. We describe the value-added components of the STAR program in this section.

STAR is intended to recognize schools that demonstrate significant gains in student achievement relative to the rest of the state, as measured by value added. STAR recognizes schools that fall within the top 15 percent of Pennsylvania schools in each grade range. All PFT-represented staff in STAR schools are eligible to receive awards for their achievement. PPS aims to recognize at least eight schools per year through the STAR program. Accordingly, if fewer than eight PPS schools place in the top 15 percent, the next-highest-ranked schools up to that number are identified in order of student growth, as long as they place in the top 25 percent of the state VAM distribution. As is the case with other school VAMs, a school's performance on STAR is based on student achievement over the two prior academic years. The first STAR schools were named for the 2011–12 school year, based on achievement results in spring 2011 and spring 2012.

Pittsburgh's collective bargaining agreement requires a statewide comparison for determination of STAR awards, so STAR VAMs included only outcomes available statewide. Because STAR requires that only statewide assessments be used, VAMs were estimated using only PSSA scores from grades 4 to 11 and Keystone algebra I and literature scores. The VAMs for STAR, based only on statewide assessments, thus differ from those used for the normal school VAM reporting. Specifically, we estimated statewide VAMs and develop for each grade range (4-5, 6-8, and 9-12) a single composite measure including all the state assessments. Schools with grade configurations of K-8 or 6-12 received composite scores that include all STAR outcomes from the relevant grades. We then used the composite VAMs to determine which schools place in the top 15 (or 25) percent of the statewide distribution. To simplify the process for calculating whether a school is in the top 15 (or 25) percent of the statewide distribution, the STAR composite is reported as a percentile rank rather than an NCE.

In Table IX.1, we show which assessments were used from 2012 and 2013 to identify STAR schools in each grade range. Note that, as decided by PPS and the PFT, the STAR composite includes the 11th-grade science PSSA, which was not included in other VAM analyses. STAR awards were determined based on the overall composite VAMs.

Table IX.1. Assessments used to determine the STAR award system by grade range, 2011–13 school years

	Elementary school grades K to 5	Middle school grades 6 to 8	High school grades 9 to 12
Overall composite	PSSA math (4,5)	PSSA math (6,7,8)	Keystone algebra I (9)**
	PSSA reading (4,5)	PSSA reading (6,7,8)	Keystone literature (10)**
	PSSA writing (5)	PSSA writing (8)	PSSA math (11)*
	PSSA science (4)	PSSA science (8)*	PSSA reading (11)*
			PSSA writing (11)*
			PSSA science (11)*

Note: The grade level of the majority of students follows each assessment in parentheses.

* Used in the 2011–12 school year only.

** Used in the 2012–13 school year only.

Table IX.2. Weights given to subjects in STAR composite by school grade range

	Elementary school grades K to 5	Middle school grades 6 to 8	High school grades 9 to 12
Math	0.37	0.46	0.40
Reading	0.40	0.39	0.43
Science	0.10	0.06	0.10
Writing	0.13	0.09	0.07

REFERENCES

Aaronson, D., L. Barrow, and W. Sander. "Teachers and Student Achievement in Chicago Public High Schools." *Journal of Labor Economics,* vol. 25, no. 1, 2007, pp. 95–135.

Buonaccorsi, J.P. *Measurement Error: Models, Methods, and Applications*. Boca Raton, FL: Chapman & Hall/CRC, 2010.

Chetty, R., J.N. Friedman, and J.E. Rockoff. "Measuring the Impacts of Teachers I: Estimating Bias in Teacher Value-Added Estimates." NBER Working Paper No. 19423. 2013.

Goldhaber, D., and D. Chaplin. "Assessing the 'Rothstein Test.' Does It Really Show Teacher Value-Added Models Are Biased?" Mathematica Policy Research Working Paper. 2012. Available at [http://mathematica-mpr.com/publications/redirect_PubsDB.asp?strSite=PDFs/education/rothstein_wp.pdf].

Hill, C.J., H.S. Bloom, A.R. Black, and M.W. Lipsey. "Empirical Benchmarks for Interpreting Effect Sizes in Research." *Child Development Perspectives,* vol. 2, no. 3, 2008, pp. 172–177.

Hock, H., and E. Isenberg. "Methods for Accounting for Co-teaching in Value-Added Models." Mathematica Policy Research Working Paper. 2012. Available at [www.mathematica-mpr.com/publications/pdfs/education/acctco-teaching_wp.pdf].

Isenberg, E., and H. Hock. "Measuring School and Teacher Value Added in DC, 2011–12 School Year." Final report submitted to the District of Columbia Public Schools. 2012. Available at [dcps.dc.gov/DCPS/Files/downloads/In-the-Classroom/IMPACT%20Guidebooks/Measuring%20Value%20Added%20in%20DC%202011-12.pdf].

Isenberg, E., and E. Walsh. "Accounting for Co-teaching: A Guide for Policymakers and Developers of Value-Added Models." Mathematica Policy Research Working Paper. 2013. Available at [http://mathematica-mpr.com/publications/redirect_PubsDB.asp?strSite=PDFs/education/accounting_for co_teaching.pdf].

Johnson, M., S. Lipscomb, B. Gill, K. Booker, and J. Bruch. "Value-Added Models for the Pittsburgh Public Schools." Report to the Pittsburgh Public Schools. Cambridge, MA: Mathematica Policy Research. 2012. Available at [http://mathematica-mpr.com/publications/redirect_PubsDB.asp?strSite=PDFs/education/value-added_pittsburgh.pdf].

Kane, T.J., and D.O. Staiger. "The Promises and Pitfalls of Using Imprecise School Accountability Measures. *Journal of Economic Perspectives,* vol. 16, no. 4, 2002, pp. 91–114.

Kane, T.J., and D. Staiger. "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation." *NBER Working Paper No. 14607,* 2008.

Kane, T.J., D.F. McCaffrey, T. Miller, and D.O. Staiger. "Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment." MET Project Research Paper, 2013.

Koedel, C., and J. Betts. "Value Added to What? How a Ceiling in the Testing Instrument Influences Value-Added Estimation." *Education Finance and Policy,* vol. 5, no. 1, 2010, pp. 54–81.

McCaffrey, D.F., T.R. Sass, J.R. Lockwood, and K. Mihaly. "The Intertemporal Variability of Teacher Effect Estimates." *Education Finance and Policy,* vol. 4, no. 4, 2009, pp. 572–606.

Meyer, Robert H. "Value-Added Indicators of School Performance: A Primer." *Economics of Education Review,* vol. 16, no. 3, 1997, pp. 283–301.

Morris, C.N. "Parametric Empirical Bayes Inference: Theory and Applications." *Journal of the American Statistical Association,* vol. 78, no. 381, 1983, pp. 47–55.

Resch, A., and E. Isenberg. "How Do Test Scores at the Floor and Ceiling Affect Value-Added Estimates?" Paper presented at the 39th Annual Conference of the Association for Education Finance and Policy. Washington, DC: Mathematica Policy Research, 2014.

Rothstein, J. "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement." *Quarterly Journal of Economics,* vol. 125, no. 1, 2010, pp. 175–214.

Rotz, D., M. Johnson, and B. Gill. "Accounting for Enrollment in After-School and Summer Programs in Teacher Value-Added Models." Memo to Pittsburgh Public Schools. Cambridge, MA: Mathematica Policy Research, 2013.

Schafer, J.L., and J.W. Graham. "Missing Data: Our View of the State of the Art." *Psychological Methods,* vol. 7, no. 2, 2002, pp. 147–177.

Schochet, P.Z., and H.S. Chiang. *Error Rates in Measuring Teacher and School Performance Based on Student Test Score Gains.* Washington, DC: U.S. Department of Education, 2010.

Tobin, J. "Estimation of Relationships for Limited Dependent Variables." *Econometrica: Journal of the Econometric Society,* vol. 26, no. 1, 1958, pp. 24–36.

Wright, S.P., J.T. White, W.L. Sanders, and J.C. Rivers. "SAS® EVAAS® Statistical Models." SAS White Paper. March 2010. Available at [www.sas.com/resources/asset/SAS-EVAAS-Statistical-Models.pdf]

www.mathematica-mpr.com

Improving public well-being by conducting high quality, objective research and surveys

PRINCETON, NJ ■ ANN ARBOR, MI ■ CAMBRIDGE, MA ■ CHICAGO, IL ■ OAKLAND, CA ■ WASHINGTON, DC

MATHEMATICA
Policy Research

Mathematica® is a registered trademark
of Mathematica Policy Research, Inc.