MATHEMATICA Policy Research

REPORT

Value-Added Models for the Pittsburgh Public Schools, 2014–15 School Year

May 3, 2016

Naihobe Gonzalez
Matthew Johnson
Brian Gill

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

Submitted by:
Mathematica Policy Research
505 14th Street, Suite 800
Oakland, CA 94612-1475
Telephone: (510) 830-3700
Facsimile: (510) 830-3701

Project Director: Matthew Johnson
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 six years that this work has been under way. Although we cannot name everyone, we especially want to thank Mary Wolfson, Tara Tucci, Veronica Amundson, Bill Hileman, Josh Aderholt, Sam Franklin, and Eddy Jones 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. Mickey McCauley ably led the process of producing individual value-added reports for teachers and schools. Mike Donaldson provided editing services. Sheena Flowers formatted the report to prepare it for public release. This report is an update to a series of previous reports on value-added models for the Pittsburgh Public Schools, and incorporates some components of these prior reports, including Johnson et al. (2012) and Rotz et al. (2014).

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

Confidence interval	A confidence interval is the range of values (for example, around a teacher value-added estimate) that the true value is expected to lie within.
Correlation coefficient	The correlation coefficient measures the extent to which two variables are linearly related, and ranges from −1 to 1. Positive correlations indicate that values of the second variable are likely to increase when values of the first variable increase. Correlations close to 0 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 value-added estimates. Smaller standard errors imply more precise estimates.
Normal curve equivalent	Each value-added measure (VAM) is reported to teachers or schools as a normal curve equivalent (NCE) ranking. NCE values range from 1 to 99, similar to percentile ranks. 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, such that a given difference in NCEs represents the same difference at any point on the scale.
R-squared	The r-squared of a model is a measure of how well the model fits the data. High values of r-squared suggest that the model is likely to predict outcomes well.
Shrinkage	Shrinkage is a procedure that helps ensure that teachers or schools with imprecise estimates are not overrepresented among high performers and low performers.
Standard deviation	A standard deviation measures how much variability from the average is in the data. Many statistics, including student test scores and teacher VAMs, tend to be distributed according to a bell curve, for which the 84th percentile is one standard deviation above average and the 98th percentile is two standard deviations above average.
Statistically significant	An estimate is statistically significant if the values in its corresponding confidence interval are either all above or all below a specified value, such as the average.

Confidence interval	A confidence interval is the range of values (for example, around a teacher value-added estimate) that the true value is expected to lie within.
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.

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

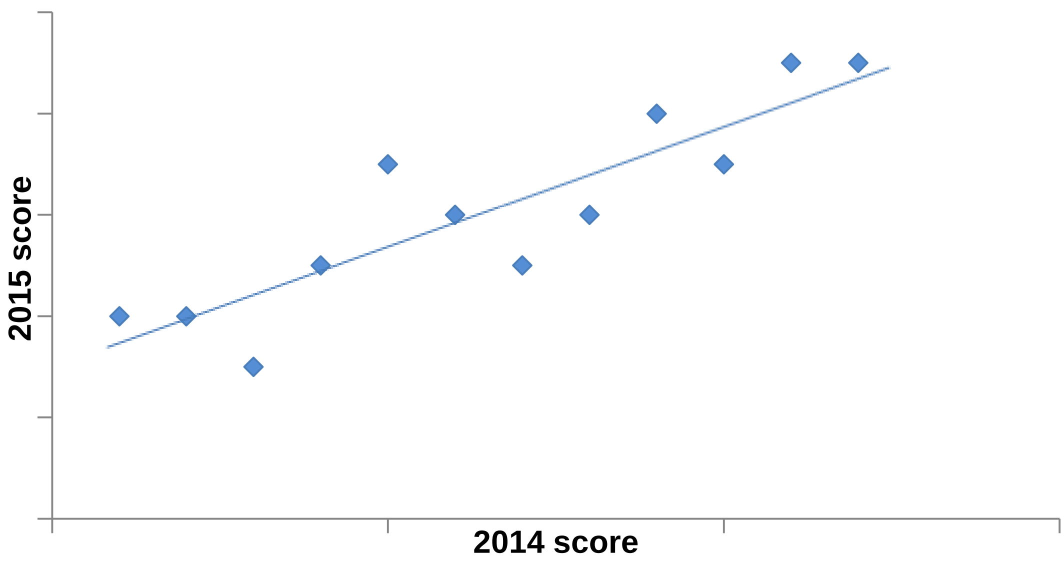
At the request of Pittsburgh Public Schools (PPS) and the Pittsburgh Federation of Teachers (PFT), Mathematica Policy Research has developed value-added models that aim to assess the contributions of individual teachers and schools to the achievement of their students. Our work in calculating value-added measures (VAMs) in PPS supports the larger, joint efforts of PPS and the PFT to empower effective teachers through evaluation, professional development, and compensation and to improve student outcomes. Students taught by high-VAM teachers are more likely to attend college and earn higher salaries (Chetty et al. 2014b). This report summarizes Mathematica's current use of VAMs to assess educational quality in PPS, updating our 2014 report on the same topic (Rotz et al. 2014).

A VAM provides a better indication of effectiveness than average test score levels or rates of student proficiency because it accounts for students' prior achievement and other factors outside the control of teachers or schools (Meyer 1997). PPS's value-added models use not only state assessments, but also course-specific assessments, student attendance, and course completion rates, with the goal of producing measures of the contributions of teachers and schools that are fair, valid, reliable, and robust. The process of measuring a teacher's or school's value added can be conceptualized as occurring in two steps.

In the first step, the statistical model uses factors such as students' prior achievement and the characteristics of students and their peers to predict their expected performance on an outcome of interest, typically a test score. These predicted outcomes represent what the students would be likely 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?

Figures I.1 and I.2 provide a simplified graphical illustration of how these predictions work, using hypothetical data. 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 2015 test score is predicted based only on the student's 2014 score in the same subject. Each pair of scores (2014 and 2015) for an individual student is represented by a diamond on the chart. The line slopes up, indicating that students with higher test scores in 2014 generally have higher test scores in 2015. Students with diamonds above the line performed better than would be predicted according to their 2014 test score, and students with diamonds below the line did not perform as well as would be predicted by their 2014 performance.

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

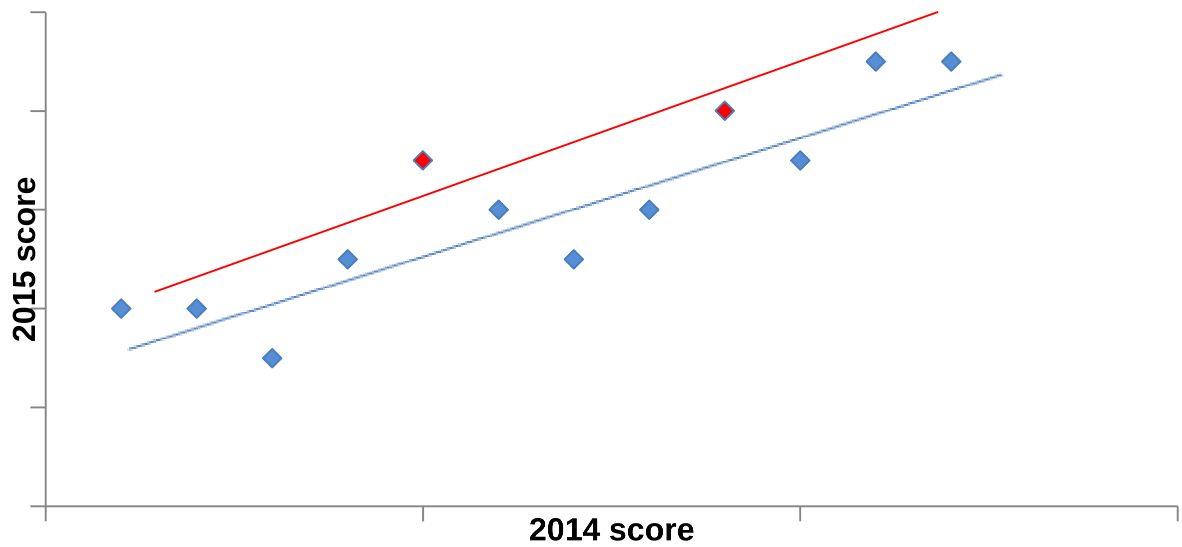


Source: Authors' visualization using hypothetical data.

Note: The diamonds in this figure depict hypothetical student test score data. The line through the points represents the prediction of students' 2015 scores based upon their 2014 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 (also using hypothetical data), 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 better in 2015 than nongifted students who had the same 2014 scores. By adjusting the prediction upward, we account for gifted status, meaning that we compare students who not only have similar baseline test scores, but are also similar in terms of gifted status. This adjusted prediction line for gifted students is represented in red.

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



Source: Authors' visualization using hypothetical data.

Note: The diamonds in this figure depict hypothetical student test score data. The red diamonds represent gifted students, and the blue diamonds nongifted students. The red and blue lines represent the prediction of students' 2015 scores based upon their 2014 scores for gifted and nongifted students, respectively.

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 have above average value added. Teachers whose classes fall short of their predicted scores have below average value added. Because VAMs are the result of a statistical model, they are estimates of the true value of a teacher's value added. Confidence intervals are reported with the VAMs and are used to determine whether a teacher's true value added is different from the average teacher's, with at least a 95 percent probability.

Value added is inherently a relative rather than absolute measure of a teacher's contribution, where the teacher is 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 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 educators and policymakers, and might vary depending on what the information is being used for. For example, when used with other measures of teacher effectiveness such as classroom observations and student survey data, the value-added threshold that triggers targeted professional development might differ from the threshold used to deny tenure.

The next three chapters 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 (also in Chapter II), discuss the technical details of the VAMs (Chapter III), and explain some limitations of the VAMs (Chapter IV).

The last five chapters 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 how VAMs are used 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 PUBLIC SCHOOLS' VALUE-ADDED MODELS

In this chapter, we describe the student outcomes and background characteristics used to calculate the VAMs for teachers and schools using PPS'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. In Section C, we describe the other variables that were included in the 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 the 2014–15 VAMs for teachers and schools. These outcomes can be placed into three general categories. The first is composed of state assessments: the Pennsylvania System School Assessment (PSSA) and the Keystone Exams. The second set of assessments includes other standardized tests administered in PPS but not given to every student in the state, such as the PSAT. The third category is made up of Pittsburgh's locally developed assessments. Table II.1 lists the assessments used in 2014–15 VAMs that PPS plans to continue using in 2015–16. We present information and results for additional assessments that contributed to the 2014–15 VAMs but will be discontinued beginning in 2015–16, such as the locally developed Curriculum-Based Assessments (CBAs), in the appendix. Table A.1 lists these assessments. If teacher VAMs were calculated for a specific grade and assessment, a T appears in the corresponding cell. Similarly, an S indicates that an assessment was used to calculate school value added.

Table II.1. Assessments used in 2014–15 values-added models that Pittsburgh Public Schools plans to continue using in 2015–16

	Grade										
Test	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
State assessments											
PSSA English Language Arts PA Core			S,T	S,T	S,T	S,T	S,T				
PSSA Math PA Core			S,T	S,T	S,T	S,T	S,T				
PSSA Science			S				S				
Keystone Algebra I							S,T	S,T			
Keystone Biology								S,T			
Keystone Literature									S		
Standardized tests not given across state											
PSAT Reading									S	S	
PSAT Writing									S	S	
PSAT Math									S	S	
Locally developed assessments											
Spanish Multimode 2									T		

Source: 2014–15 Pittsburgh Public Schools value-added models.

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.

Most of the exams listed in Tables II.1 were used in both 2014–15 teacher and school VAMs, with some exceptions. The Spanish Multimode exam was excluded from school value-added calculations because some schools have only one teacher for this subject; reporting school-level results could therefore implicitly identify a single teacher's VAM. Some assessments, such as the PSAT and the science PSSA, were used only for schools because PPS and the PFT determined that those assessments are not directly aligned with specific courses, and thus student achievement on them cannot be attributed to specific teachers. In the case of Keystone Literature, we found that this assessment did not provide information that allowed the VAMs to differentiate among teachers, and therefore it was used only for school VAMs. We will reevaluate its suitability for teacher VAMs in future years.

Whenever possible, VAMs for schools are averaged across the last two years of instruction, and VAMs for teachers are averaged across the last three. Due to changes in assessments and curricular standards over these periods, some of the assessments listed in Table A.1 did not contribute data from the 2014–15 school year (see the appendix). The most significant change during this period was the introduction of the Pennsylvania (PA) Core Standards and the resulting redesign of the math, reading, and writing PSSAs during the 2014–15 school year. The PA Core Standards set new, rigorous academic expectations in English language arts (ELA) and math for what students should master by the end of each grade level to be successful in college and careers. To ensure that the PSSAs aligned with these new standards, the state designed two new assessments in math and ELA, replacing the reading and writing exams with a single ELA assessment in each grade. The average and standard deviation of PPS students' PSSA scaled scores in 2013–14 and 2014–15 are presented in Table II.2 for each subject and grade.

As 2014–15 was the first school year the new standards were implemented, PPS (like many other districts) experienced marked changes in measured student performance. As the averages in Table II.2 show, ELA/reading and math scores fell across grade levels in PPS—though this decrease could be due to the changes in the assessments rather than changes in student achievement. In addition to receiving lower scaled scores on average, students tended to perform within a narrower range of scores in 2014–15, as reflected in the smaller standard deviations. For comparison, the average and standard deviation of the grade 4 and grade 8 science PSSAs, which did not experience a curricular realignment, remained relatively stable. Due to the magnitude of these changes for math and ELA, we recommended estimating separate VAMs for the old and new versions of the assessments. The relationship between a student's prior test scores, demographic characteristics, and most recent PSSA score was likely different under the legacy standards and new standards. Along with VAM results, we report average predicted and actual PSSA scores for each teacher to assist in the interpretation of the VAMs. Because the scale scores changed considerably, we determined that reporting students' actual and predicted scores combined over three years would be much less meaningful to teachers. As a result, we also displayed this information separately for the old and new versions of the PSSA on teacher VAM reports.

Table II.2. PSSA performance by subject and grade, 2013–14 and 2014–15

Test	Average scaled score		Standard deviation	
	2013–14 Legacy standards	2014–15 PA Core standards	2013–14 Legacy standards	2014–15 PA Core standards
ELA/Reading Grade 3	1,270	997	158	101
ELA/Reading Grade 4	1,268	982	236	110
ELA/Reading Grade 5	1,232	982	229	108
ELA/Reading Grade 6	1,253	974	245	112
ELA/Reading Grade 7	1,324	974	237	106
ELA/Reading Grade 8	1,407	982	270	103
Math Grade 3	1,258	977	194	116
Math Grade 4	1,354	959	259	101
Math Grade 5	1,359	942	257	107
Math Grade 6	1,382	922	275	95
Math Grade 7	1,404	921	255	92
Math Grade 8	1,357	918	251	92
Science Grade 4	1,323	1317	197	186
Science Grade 8	1,205	1213	205	202

Source: Authors' calculations based on data provided by Pittsburgh Public Schools.

Next year, additional changes to the assessments that are included in teacher and school VAMs will take place. Some of these changes will occur simply because another year has elapsed. For example, the grade 2 TerraNova and grade 8 physical science CBAs listed in Table A.1 will have been discontinued for more than three years and thus will not enter any VAMs. In addition, due to the elimination of the grade 2 TerraNova exams, no assessments will be available to predict student performance in grade 3. As result, the grade 3 PSSAs will no longer be used as student outcomes. Other changes will result from policy decisions by PPS. Because test security procedures were relaxed for CBAs in an effort to allow teachers to use them for formative purposes, PPS made a decision not to use these assessments as outcomes in VAM calculations in future years.

The validity of using a particular assessment in VAMs depends, first of all, on its validity as a measure 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 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 PSAT, in contrast, was not designed to align with any particular course, but was developed and refined by psychometric experts (College Board 2016).[1]

[1] Beginning in fall 2015, the College Board redesigned the total testing time, components, and scoring of the PSAT to focus on the knowledge and skills most important for college and career readiness and success. Because of the timing of the exams, the PSAT value-added models described in this report include only one year of data from fall 2014 and were thus unaffected by this change. Student performance on the PSAT and its relationship to baseline controls may change in future years.

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. Detailed results (in Chapter VII) show that almost all the VAMs were able to distinguish some schools and teachers from average in 2014–15.

Table II.3 lists the outcome measures used in VAMs along with the prior test-score measures that we used as baseline controls to account for students' prior achievement. As before, we present this information separately for discontinued assessments in the appendix (see Table A.2). 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) to maximize the predictive power of the model and (2) to minimize the number of students excluded from the model because of missing baseline test scores.

Whenever possible, we use 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 can provide additional information about students' baseline knowledge and abilities. Thus, all models include at least two tests from prior years. We also account for a third prior test in all cases in which adding the third prior test could 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 use only baseline tests that most students took in the year prior to the current test. The one exception to this rule is grade 8 PSSA scores, which are generally available for high school students and are thus included as baseline scores in value-added models for 10th-, 11th-, and 12th-grade outcomes.

Table II.3. Assessment outcomes and baseline test scores

Outcome	Prior test 1	Prior test 2	Prior test 3
PSSA Math PA Core Grade 4	PSSA Math Grade 3	PSSA Reading Grade 3	
PSSA ELA PA Core Grade 4	PSSA Reading Grade 3	PSSA Math Grade 3	
PSSA Science Grade 4*	PSSA Math Grade 3	PSSA Reading Grade 3	
PSSA Math PA Core Grade 5	PSSA Math Grade 4	PSSA Reading Grade 4	PSSA Science Grade 4
PSSA ELA PA Core Grade 5	PSSA Reading Grade 4	PSSA Math Grade 4	PSSA Science Grade 4
PSSA Math PA Core Grade 6	PSSA Math Grade 5	PSSA Reading Grade 5	PSSA Writing Grade 5
PSSA ELA PA Core Grade 6	PSSA Reading Grade 5	PSSA Writing Grade 5	PSSA Math Grade 5
PSSA Math PA Core Grade 7	PSSA Math Grade 6	PSSA Reading Grade 6	
PSSA ELA PA Core Grade 7	PSSA Reading Grade 6	PSSA Math Grade 6	
Keystone Algebra I Grade 8	PSSA Math Grade 7	PSSA Reading Grade 7	
PSSA Math PA Core Grade 8	PSSA Math Grade 7	PSSA Reading Grade 7	
PSSA ELA PA Core Grade 8	PSSA Reading Grade 7	PSSA Math Grade 7	
PSSA Science Grade 8*	PSSA Math Grade 7	PSSA Reading Grade 7	

[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.

Outcome	Prior test 1	Prior test 2	Prior test 3
Keystone Algebra I Grade 9	PSSA Math Grade 8	PSSA Reading Grade 8	PSSA Writing Grade 8
Keystone Biology Grade 9	PPSA Science Grade 8	PSSA Math Grade 8	PSSA Reading Grade 8
Keystone Literature Grade 10*	CBA ELA I Grade 9	PSSA Reading Grade 8	PSSA Writing Grade 8
Spanish Multimode Level 2**	Spanish 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
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

Source: 2014–15 Pittsburgh Public Schools value-added models.

Note: Grade level refers to the grade most students are in when taking the exam.

*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 PPS are 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 are 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 differs slightly from the method used for test-based VAMs; see Chapter III for details.

Table II.4 lists the non-test outcomes and their baseline measures. We include baseline test measures alongside baseline measures of the outcome of interest because the test measures typically improve the predictive power of the model (that is, current attendance rate and core pass rate are related to previous achievement as well as to 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. For these VAMs to be valid, PPS must ensure that the standards for passing a core course or being marked absent are consistent across schools.

Table II.4. Non-assessment outcomes and baseline measures used in school VAMs

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

Source: 2014–15 Pittsburgh Public Schools value-added models.

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 VAM 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 PPS's course enrollment and grades data, we determine the percentage of core courses that a student passes, and then apply a value-added statistical model to that percentage (that is, we assess whether the percentage is 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 PPS high schools: 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 differences.

2. Attendance rate

For grades 4 through 12, we calculate 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 suggests that standards for determining whether an absence is excused or unexcused may vary over time and among schools. We therefore use the overall attendance rate in the VAMs, which is more stable over time and across schools than excused or unexcused absences. As with other VAMs, the attendance VAM does not use the raw attendance rate as the measure of school performance, but rather measures the extent to which the school's students are attending at higher or lower rates than predicted, given their attendance rates in the preceding year and other baseline characteristics, including whether they were only enrolled in PPS for part of the year.

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

Each VAM accounts for observable student characteristics to help isolate the effect of teachers and schools on student achievement. The factors that are 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.5 defines the student background characteristics that are included in teacher and school VAMs.

In addition to characteristics of individual students, we account for classroom average characteristics in most of the measures.[3] These measures of a student's peers account for classroom composition based on gender, meals program eligibility, English-language-learner status, gifted status, disability status, prior-year absence rate, prior-year suspension rate, prior-year full-year district membership, and prior-year average PSSA math and reading scores.[4] These classroom characteristics are created by averaging individual characteristics across

[3] 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.

[4] At the high school level, the classroom averages for prior PSSA math and reading scores came from grade 8, regardless of the high school grade.

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 to capture the totality of the student's experience with that subject.

Teacher VAMs also account 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. Average class size is not included as a control 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.[5] Because course selections are made at the beginning of the school year, they are outside the control of current-year teachers. Thus, we account for students' enrollment in these programs to protect the teacher VAMs from being biased due to differences in the curricula taught in advanced courses. We classify students as being enrolled in these programs on a yearly basis, pooling information across courses for each subject. However, we omit 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 receive data from PPS indicating whether students ever entered a magnet school lottery (regardless of whether the application was successful). We include this variable in both school and teacher VAMs 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 explain most of the variation in students' current test scores. However, some student-level characteristics tend to be statistically significant as well, even when 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 are related to student test scores. Beginning in 2015–16, free and reduced-price lunch status will be replaced by a new indicator of students' meal eligibility. After transitioning to Community Eligibility Provision, which provides free lunches to all students in the district, PPS no longer collects free and reduced-price lunch status data. Under the Community Eligibility Provision program, meal status is collected directly from the Department of Public Welfare.

Table II.5. Variables for student background characteristics in teacher and school value-added models, 2014–15

Background variable	Definition
Male	Male gender
Meals program*	Free or reduced-price lunch 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)

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

Background variable	Definition
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 the current school year
Repeated assessment***	Took the same outcome assessment in a previous year
Modified exam	Student took an alternate version of the exam
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

Source: 2014–15 Pittsburgh Public Schools value-added models.

Note: 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.

*Beginning in the 2015–16 school year, eligibility for free or reduced-price meals will be replaced by an indicator of whether a student is economically disadvantaged.

**Indicates that the variable is excluded from school value-added models.

***This variable was not used in estimating PSAT VAMs because students often take the PSAT in more than one year. Instead, an indicator of whether the student repeated the current grade is used.

III. TECHNICAL DETAILS OF PPS VALUE-ADDED MODELS

This chapter explores the technical details of the value-added statistical models calculated for PPS. We first detail the models and then discuss how we prepare the data for analysis and process the output to calculate the final VAMs reported to teachers and schools.

A. Detailed value-added model description

The following general statistical equation describes the value-added models[6]:

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

In the model, $Y_{i,t,c}$ is the outcome for student *i* in year *t* and classroom *c*. The models are 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 2014–15. $B_{i,t-1}$ is a vector of baseline scores for student *i* from a prior year to account for students' own academic histories. The relationship between these baseline scores and the outcome is estimated separately for each year in the model, to account for minor changes in the assessments that could have occurred over time. The baseline scores typically come 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 are used as control variables in all high school VAMs. We include 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 comes from the same subject area as the outcome measure.[7] See Table II.3 and Table A.2 for the full list of outcome measures and baseline variables.

X_i is a set of variables for observable student characteristics, whereas $\bar{X}_{i,c}$ represents average observable peer characteristics (described in Section C of Chapter II).[8] D_i is a set of variables for a student's teachers in the subject of interest or a student's 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 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.

---

[6] This model is a fixed-effects dynamic ordinary least squares model, which has been shown in simulations to be robust to a variety of assumptions about the student assignment process (Guarino et al. 2014).

[7] An example of an exception is the grade 4 science PSSA, where a same-subject baseline score is 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 and $\bar{X}_{i,c}$ 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, this typically improves the validity of VAM estimates as long as both the related variables are relevant to student achievement growth. (Note that we identify the relationship between student and classroom characteristics using variation within teachers and schools.)

We define the average VAM (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. Rather, it means that positive VAMs represent above-average (above predicted) teacher or school performance and that negative VAMs represent below-average (below predicted) teacher or school performance. For test-based school VAMs, average performance is defined at the state level by creating a hypothetical distribution of statewide performance in a separate step, as described in Chapter VI. For attendance and core course pass rate school VAMs and for all teacher VAMs, average performance is defined among PPS schools and 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 2013–14 and 2014–15 school years in this report. Models for teachers examine up to three years of teaching, producing an average of the teacher's performance across the 2012–13, 2013–14, and 2014–15 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 use multiyear VAMs when possible because, compared to single-year measures, 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 high-stakes applications (see McCaffrey et al. 2009). However, multiyear VAMs are less reflective of immediate past performance, because they average annual value-added scores over multiple years.

B. Standardization

Because VAMs 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 of scores. Expressing scores this way allows us to interpret above-average scores in terms of how close to average most students tend to fall, regardless of the assessment. When calculating VAMs, we mean-center the data for all baseline and 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 are calculated relative to the contribution of the average teacher or school in the sample.

C. Accounting for measurement error

Test scores are imperfect measures of student ability. To account for this imprecision, we use a statistical technique known as an errors-in-variables measurement error correction (Buonaccorsi 2010). The errors-in-variables correction helps alleviate problems related to measurement error in pre-tests by incorporating information about the test-retest reliability of

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

those tests.[10] We use grade- and subject-specific test reliability data available from test publishers' websites.[11] Most VAMs account for prior achievement on tests for which reliability information is available (specifically, the PSSA and PSAT exams). In 2014–15, the reliability of these exams ranged between 0.81 and 0.95. In some VAMs, we also include controls for prior-year achievement on CBAs, TerraNova, or Multimode exams, for which reliability information is unavailable. To account for measurement error in these instances, we use the reliability of the PSSA in the same subject and year and in the nearest grade. Although using the reliability information of another test is not ideal, it is preferable to assuming that these exams are perfectly reliable.

We implement the measurement error correction using the two-step estimation procedure described in Isenberg and Hock (2012). In the first step, we use an errors-in-variables regression to obtain estimates of the pre-test coefficients for each outcome assessment. We then use these measurement-error-corrected coefficient values to calculate the adjusted test score gain for each student, net of the contribution attributable to the student's baseline performance. In the second step, this net test score gain is regressed on the remaining control variables and the indicators for students' teachers in the subject of interest or students' schools during the year. This last regression yields the VAMs.

D. Teacher and school dosage

When a student was in a teacher's class or in a school for only part of the school year, we use a dosage approach to account for the amount of time that teacher or school had to influence the achievement growth of this student compared to that of students who were in the teacher's class or in the school for the full year. Dosage can be thought of as a weight that is applied to each student when calculating a teacher or school's value added. We use the "Full Roster Method" to account for differences in dosage when calculating value added (Hock and Isenberg 2012; Isenberg and Walsh 2015).

School dosage is based on the share of school days a student was enrolled in a VAM-eligible course at a school, which we determine using school enrollment and course roster data provided by PPS. For example, if a student moved schools in Pittsburgh during the year and was enrolled in a math class at each school, the school dosage values in the relevant math assessment VAMs are fractions between 0 and 1 based on the days the student was enrolled at each school.[12] In the case of the PSAT math, reading, and writing exams, which are not attributed to particular courses, students had to be enrolled in a core course in the same subject as the relevant PSAT assessment to be included in the PSAT VAMs. Similarly, because non-test-based outcomes cannot be attributed to particular courses, students are only required to have enrolled in at least one core course to be included in the attendance and core pass rate VAMs. The school dosage for

[10] 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.

[11] PSSA reliability was obtained from the Pennsylvania Department of Education technical reports. PSAT reliability data was obtained from the College Board.

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

these outcomes is determined by the share of school days the student was enrolled at the school. This core course enrollment restriction avoids potential errors in administrative data and ensures that students are fully attached to the schools to which they are attributed.

To obtain accurate dosage for teachers, PPS provides roster verification data that is also used by the state for the Pennsylvania Value Added Assessment System (PVAAS). Teachers verify that they are linked accurately to students for each assessment. Specifically, teachers confirm the percentage of days a student was enrolled in instruction with them and the percentage of content-specific instruction they were responsible for providing to that student for each assessment. For example, if a student was enrolled in a teacher's class during the entire school year but left class to receive pull-out instruction one day out of a five-day school week, then the student would have an 80 percent dosage with that teacher. The roster data verified by teachers are then approved by the school principal and reviewed by staff in the central office.

Only teachers who teach a VAM-attributed course in a given assessment are eligible to receive a VAM for that assessment. For example, only teachers who teach math courses explicitly tied to the CBA geometry grade 10 exam receive VAMs for this assessment. Teachers who teach other 10th-grade math courses do not receive CBA geometry VAMs, even if their students happen to take the CBA geometry exam. Although the PVAAS roster file shows which students were taught by which teachers for each VAM-eligible assessment, it does not contain information on exactly which courses students took. Therefore, we also employ the course roster file provided by PPS, which contains data on all courses taken by students according to the district's administrative records.

There is a lag between when the PVAAS roster confirmation process takes place and when final decisions are made about which courses should be linked to VAM-eligible assessments. To prevent students from being attributed to teachers when they were enrolled in courses that are not linked to VAM-eligible assessments, we use the course roster file to confirm whether students listed in the PVAAS roster file took VAM-eligible courses with the correct teachers. A student is therefore attributed to a teacher for the purposes of estimating a VAM for a particular assessment only if (1) the student is attributed to that teacher in the PVAAS roster file for that assessment and (2) the student is listed as taking a VAM-eligible course tied to that assessment with that teacher in the course roster file.

Students who are listed as taking a VAM-eligible course in the course roster file but who do not appear in the PVAAS roster file are still included in the VAM, but these students are not attributed to any teacher receiving a VAM report. Rather, we assign these students to a generic school-specific record so we can keep them in the analysis. The model treats the students as being taught by a generic teacher at their school, even though they may have received instruction for that outcome from multiple teachers. The benefit of including these students in a catch-all, school-specific record is that they contribute to the accuracy of the model by increasing the precision of the estimated coefficients on the control variables. Because VAMs with too few students are highly imprecise, we also assign students taught by teachers with fewer than five student equivalents to the school-specific record.[13] On average across all assessment-level VAMs, 9.2 percent of student-teacher records were treated this way in 2014–15.

Teacher dosage for students appearing only in the course roster file is based on school and course enrollment data. If a student appears in the PVAAS roster as taking a VAM-eligible assessment but this student is not enrolled in a VAM-eligible course tied to that assessment according to the course roster file, then this student is excluded from the model. The student is excluded because the course roster file should contain the universe of students and course enrollments, which means that cases where a student-assessment combination appears in the PVAAS roster file but the corresponding student-course combination does not appear in the course roster file are likely the result of erroneous data in the PVAAS roster file.

E. Shrinkage

We use a procedure known as empirical Bayes estimation, or shrinkage, to address the fact that among teachers or schools with the same level of true performance, those with fewer students 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 VAMs—will tend to be overrepresented at both the high and low ends of the estimated performance distribution just by chance.

The shrinkage adjustment we use is an empirical Bayes procedure based on Morris (1983), which accounts for the fact that a result with greater precision carries greater strength of information about a teacher's true performance level.[14] The adjusted result is a weighted average of the individual's initial result and the mean result across teachers, with more precise initial results 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 VAMs, re-centering them around zero before and after shrinkage.

To further minimize the risk of making erroneous conclusions on the basis of imprecise measures, we report the VAMs only of teachers who taught more than 10 student-equivalent observations across the three years of data included in the model. This type of reporting restriction, commonly used in research, reduces the potential that teacher effects influenced by the scores of just one or two students are misinterpreted (Kane and Staiger 2002; McCaffrey et al. 2009). We impose a similar restriction in school models, but the restriction is generally not binding because most schools have many more than 10 student equivalents taking each assessment.

[13] Each student is weighted by his or her dosage in calculating whether a teacher taught less than 5 student-equivalents. The term *student equivalent* reflects that each student record is weighted by the corresponding dosage.

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

F. Normal curve equivalent reporting units

Each VAM is reported to teachers or schools as a normal curve equivalent (NCE) ranking.[15] NCEs are similar to percentile ranks in that they are reported on a 99-point scale with an average of 50. An NCE is equivalent to a percentile rank at scores of 1, 50, and 99.[16] 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.

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 measures where they stand in the distribution of schools serving the same grades *across Pennsylvania*. Ideally, we would use a statewide comparison for all VAMs, but many teachers have VAMs based on student assessments conducted only in PPS, precluding a statewide comparison.

G. Accounting for very imprecise VAMs

All measures of teacher effectiveness (including observation-based measures of teacher performance as well as value added) 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 due to factors unrelated to a teacher's actual value added (such as whether a student was ill when he took the test, leading him to perform worse than he typically would). This is sometimes called statistical noise. Precision is reduced by statistical noise and is increased when more students take an assessment, because there is more information available to use in measuring performance.

In rare cases, a teacher may have more than 10 student equivalents contributing to a VAM but receive a very imprecise VAM, reflected by a wide confidence interval. In the extreme case when the confidence interval of a teacher's VAM is so wide that it spans the full range of NCE values, the VAM is too imprecise to provide valuable information about the teacher's true effectiveness. For this reason, we do not report VAMs that span the complete NCE range (1–99) and do not include these results in composite calculations. This exclusion affected only one teacher in 2014–15.

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

It is possible for a teacher's VAM to change from year to year even if the same set of students are attributed to that teacher. This occurs because the set of students taking relevant assessments during the most recent year is added to the current year's models, and the set of students from four years ago is omitted. This difference in the sample of students entering the model results in changes to the estimated coefficients on all variables, as well as to the comparison group of teachers in the VAMs. To avoid introducing variation in the reported

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

[16] NCEs below 1 and above 99 are bottom- and top-coded at these values for reporting purposes.

VAMs over time that is unrelated to actual changes in teacher effectiveness, PPS decided not to provide new VAMs to teachers with the exact same set of students in the value-added models as the preceding year. Although these teachers are included in the current year's model to improve precision, they receive the same VAMs in their reports as they did the preceding year. Among all assessment-level teacher VAMs, 23 percent were replaced with a previous year's value in 2014–15.

I. Technical details of VAMs for non-test outcomes

Because many students achieve perfect attendance and 100 percent core pass rates, the maximum possible values, the statistical model for these two outcomes differs slightly from the primary model described by Equation (1) at the beginning of the chapter. For both of these VAMs, we use a Tobit version of Equation (1). The Tobit model separately calculates 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). In addition, we include grade-level indicators to account for the possibility that core pass rates and attendance rates vary systematically across grade levels.

- **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 such situations, ordinary linear regression models like Equation (1) provide biased results, because the relationship between higher values for the background variables and a higher core course pass rate becomes nonlinear. To account for this bias, we use a Tobit model. In the core pass rate VAMs, we account 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 passed all core courses in the prior year, we include an indicator variable for whether 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 also calculate this VAM using the Tobit model. We include 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 in the model. When we attempted to calculate attendance rate VAMs using data in grades K–3, we found that the results were very imprecise due to relatively small variation in attendance rates among students in these grades. We therefore limit 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 value-added models used in PPS.

A. Nonrandom assignment of students

Because students are not randomly assigned to teachers, it is possible that the assignment process can lead to bias in VAMs (Rothstein 2010). Students with characteristics that are not observed in data but that can affect test score gains, such as having involved parents or high levels of motivation, may be assigned to particular teachers, leading to biased results of teachers' true value added. The value-added models used for PPS assume that assignment is unrelated to unobservable characteristics, or as good as random, once we have accounted for a large set of characteristics available in the data.

Even though this assumption may not be strictly true, research suggests that the resulting bias in the VAMs is likely to be small. Goldhaber and Chaplin (2015) found that sorting bias is small relative to the variability of teacher VAMs. Kane and Staiger (2008) and Kane et al. (2013) found that variation in teacher VAMs 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. (2014a) suggests the finding holds even when assignment is not random. The authors analyzed the average test scores of students before and after teachers left or entered their schools. They found that students' learning responded as expected to changes in their teacher's value added. For example, after a high-value-added teacher changed schools, the average test scores of the grade from which she left decreased and the average test scores of the grade into which she transferred increased. Work by Bacher-Hicks et al. (2014) in a different context validated this approach and obtained consistent findings. This body of research suggests that a teacher's VAM reflects 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 is likely small and does not prevent VAMs from identifying an important component of teacher performance.

Nevertheless, there are some important characteristics that perhaps should be included in the models but are not available. For example, PPS's data system does not currently track student participation in all programs (during the school year or 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 available but is not included in the statistical models. We examined the suitability of including control variables for these programs in an earlier analysis (Rotz et al. 2013) and determined that it is possible for teachers to play a role in student enrollment in many after-school programs. Thus, controlling for enrollment in after-school programs could actually lead to greater bias in VAMs. Participation in the Summer Dreamers Academy occurs in the summer before a student is taught by a teacher and therefore cannot be influenced by that teacher. However, when we experimented with adding a control for Summer Dreamers Academy enrollment, we found that teacher VAMs were virtually unchanged, with correlations between the VAMs in both models above 0.99. Thus, rather than making an exception for Summer Dreamers, we recommended that our statistical model not account for any extracurricular activities.

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 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. As in other studies of value added, however, we find that after accounting for students' baseline performance, additional control variables have little effect on teachers' VAMs. Together with past research that documents limited sorting bias once baseline performance is accounted for, this suggests that characteristics that are unobserved are likely to have a minimal impact on the results.

B. Distinguishing between school and teacher effects

When we measure 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 also likely to remove some of the teacher-specific performance. To avoid these problems, we do not include school indicators in the teacher VAMs.

An alternative method to account for the influence of schools in teacher 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, available resources, principal quality, and school safety), because data are not readily available on those characteristics. Even if these data were 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 PPS data, such as schoolwide averages of the number of days students are suspended, the 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, teacher VAMs 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 after the other control variables have been taken into account 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 VAMs 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 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 the following year's VAMs differently from the way it would if the school had allowed the student to progress to the next grade.[17]

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, because no prekindergarten measure of achievement exists. We therefore conducted a sensitivity analysis and found that using prior-year test scores produced results similar to using pre-entry test scores at the middle and high school levels (Lipscomb et al. 2010). Because baselines from the prior 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 the prior year for both school and teacher VAMs.

D. Missing data

Students may be missing background data that are required by the value-added model for a number of reasons.

1. Missing baseline test scores

In each value-added model some students who have data on the outcome measure are dropped because of missing data on at least one baseline or background variable. In most cases, the missing element is a prior test score.[18] Prior scores could be missing for several reasons, such as if students transferred into PPS from outside the district, took a test out of grade, or were absent from school during testing in the prior year.

To avoid excluding these students from the analysis, 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. In addition, imputation is not feasible

[17] 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.

[18] Across all teacher value-added models, the median percentage of students excluded from the sample due to missing data was 13 percent in 2014–15.

for students who transferred to PPS from other districts, because there is no information in PPS's data collection on their prior scores. Because a poorly imputed variable may cause a loss in precision of the coefficient on prior test scores and ultimately of VAMs, we do 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 2013–14 and was missing prior test score data, he or she would have taken the normal end-of-year assessments in PPS. That student would have been dropped from the 2013–14 VAMs due to missing prior test score data, but would appear in the 2014–15 VAMs because the end-of-year assessments in 2013–14 could be used as the baseline scores. Therefore, except in cases of rapid mobility, students tended to be included in the VAMs after a full year in the district. Although missing baseline scores are not problematic in most cases, high-school entry is a special case. A student missing 8th-grade PSSA scores is permanently excluded from all the high school VAMs.

2. Missing data on classroom average characteristics

In some rare cases, we have data available for a student but not for his or her peers. This typically happens when we cannot assign a student to a particular classroom. Because we include controls for average classroom characteristics in the value-added models, these students would be omitted from the analysis unless their characteristics were imputed. We use the simplest imputation possible for these average classroom 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. In 2014–15, this imputation was applied to no more than 0.3 percent of student-teacher records across all VAMs. By making this assumption, we avoid dropping records without adding additional complexity to the model.

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 VAMs. The concern is based on the premise that because value added is a measure of growth over the course of a 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 would appear to have nowhere to go but up (a floor effect).

It is unlikely that these effects lead 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 VAMs. 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 VAMs (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 only a concern when a large proportion of students have scores at the maximum value. In the assessments used by PPS in the 2014–15 school year, about 0.6 percent of student test scores were at the maximum and 1.2 percent were at the minimum. Table IV.1 further breaks down the students who scored at the maximum or minimum observed value, by test and subject, pooling across grades. For almost all assessments, maximum and minimum scores are very rare. The largest percentage of students scored at the maximum on the Spanish Multimode exams. Minimum scores are most common on the PSAT and PSSA Science exams, in both 4th and 8th grades. 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 and minimum values, by test type and subject, 2014–15

Test	Students taking assessment	Students scoring at minimum value		Students scoring at maximum value	
		Number	Share	Number	Share
PSSA ELA PA Core Exams	9,869	10	0.001	6	0.001
PSSA Math PA Core Exams	9,961	9	0.001	7	0.001
PSSA Science Exams	3,271	245	0.075	4	0.001
CBA Math Exams	6,304	22	0.003	145	0.023
CBA ELA Exams	6,615	9	0.001	18	0.003
CBA Science Exams	6,185	11	0.002	35	0.006
CBA Social Studies Exams	5,553	9	0.002	31	0.006
Keystone Algebra I	3,301	4	0.001	5	0.002
Keystone Literature	1,841	4	0.002	13	0.007
Keystone Biology	2,024	110	0.054	7	0.003
PSAT Math	2,745	84	0.031	10	0.004
PSAT Reading	2,747	125	0.046	7	0.003
PSAT Writing	2,585	124	0.048	1	< 0.001
Spanish Multimode	1,595	3	0.002	81	0.051

Source: Authors' calculations based on data provided by Pittsburgh Public Schools. Results are based on outcome assessments from 2014–15.

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, which can be problematic if the tests do not accurately capture the content and skills emphasized by the curriculum, or by 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, 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 the 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 is therefore 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 PPS has VAMs based on several different outcomes, with assessments spanning multiple grades and covering a variety of subjects. Many teachers likewise have VAMs related to more than one assessment. At the policy direction of PPS and the PFT, the VAMs for the individual test-based outcomes are aggregated into composites 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 subject-wide and overall performance. Composite VAMs are calculated separately by subject (for schools only) and across all test-based measures (for schools and teachers). For example, the math composite for an elementary school incorporates VAMs from the math PSSAs in grades 3, 4, and 5.

A. Composition of composite VAMs

Table V.1 shows how all the individual assessments were grouped into school composite VAMs in 2014–15. Beginning in 2015–16, these composite VAMs will only include the assessments listed in Table II.1. For example, the math composite VAM for an elementary school will include math PSSAs only in grades 4 and 5. Composite VAMs are calculated based on the assessments available in the grade ranges taught at a school. Schools with grade configurations of K–8 or 6–12 receive composite VAMs that include all assessments from the relevant grades. PPS schools are then ranked together based on their effectiveness rating for each composite VAM. STAR awards use a different composite, described in Chapter VIII.

Table V.1. The composition of subject composite VAMs for Pittsburgh Public Schools, 2014–15

	Elementary school grades K to 5	Middle school grades 6 to 8	High school grades 9 to 12
Math composite	PSSA math PA Core (3, 4, 5) PSSA math (3,4,5)	PSSA math PA Core (6, 7, 8) PSSA math (6,7,8) CBA math (6,7,8) CBA Algebra I (8) Keystone algebra I (8)	CBA algebra I (9) Keystone algebra I (9) CBA geometry (10) PSAT math (10, 11) CBA algebra II (11)
English/language arts composite	PSSA ELA PA Core (3, 4, 5) PSSA reading (3,4,5) PSSA writing (5)	PSSA ELA PA Core (6, 7, 8) 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) CBA ELA III (11) CBA ELA IV/African American literature (12)
Science composite	PSSA science (4)	CBA earth science (6) CBA life science (8) CBA physical science (7,8) PSSA science (8)	CBA biology (9) Keystone biology (9) CBA chemistry (10) CBA physics (11)
Social studies composite	n/a	CBA world geography (6) CBA world history (7) CBA US history (8)	CBA civics (9) CBA world history (10) CBA US history (11)

Source: 2014–15 Pittsburgh Public Schools value-added models.

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

n/a = not applicable.

Although we calculate subject composite VAMs for all grades, as shown in Table V.1, not all are reported because some schools have only one teacher for certain subjects. Specifically, 4th grade PSSA science is not reported in subject composite VAMs for schools with grades K–5, and middle school social studies CBAs are not reported in subject composite VAMs for schools with grades 6–8. Reporting results for subjects taught by a single teacher would implicitly identify that teacher. Science composites are reported for K–8 schools, and science and social studies composites are reported for 6–12 schools, where they are combined with other VAMs (and other teachers). They are also included in the overall test-based composite (averaged across subjects) for schools in all relevant grade configurations.

Spanish Multimode VAMs are similarly excluded from school VAMs, as some schools have only one Spanish teacher. In addition, as is described in more detail in Chapter VI, the school composite VAMs 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.

In addition to the VAMs for each assessment, teachers receive an overall composite that combines the information in the individual VAMs and compares teachers' performance to all other teachers in the district with VAMs. Teachers with just one VAM-eligible assessment can receive a composite VAM that is different from the individual VAM for that assessment because the comparison group for the composite includes all teachers with VAMs in the district, and not only teachers who taught the same assessment.

B. Construction of composite VAMs using weights

The composite measures are obtained by combining the assessment-level VAMs of each teacher or school using a multistep method. We first calculate a weight for each VAM based on the precision with which the model was estimated and the number of students contributing to the VAM for each teacher. We then place all assessment-level VAMs on the same scale and, for each teacher or school, average these components together using the weights calculated in the first step.[19]

In the first step, we calculate precision weights for each assessment-level VAM using a methodology approved by PPS and the PFT. Here, precision refers to the average variance of the VAMs from each model. VAMs tend to be estimated less precisely when fewer students are included in the models or when the assessments being used provide less accurate measures of student achievement. The advantage of precision weighting is that more weight is given to components that are estimated more precisely, which thereby produces a composite VAM with less statistical noise than alternative weighting schemes. However, a drawback is that it 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 VAMs may not be the

[19] The standard errors of the composite scores are calculated using the fact that the variance of a sum is a function of the variances of each element and their covariances. We estimate 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 is most important for long-term success, or the 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 an overall composite VAM that has more statistical noise than a precision-weighted composite but might better align with PPS's educational goals.

Precision weights are created based on the average precision of teacher or school VAMs for a given assessment. This implies that without further adjustments, the same weights would be used in creating the composites for teachers receiving individual VAMs for the same assessments. For example, suppose Teacher A taught 4th-grade math to 60 students and 5th-grade math to 20 students, and Teacher B taught 4th-grade math to 20 students and 5th-grade math to 60. With weights developed based only on the average precision of VAMs, these teachers' 4th- and 5th-grade math VAMs would receive the same weight in their composite VAMs. This may be undesirable, as Teacher A teaches mostly 4th-grade students and Teacher B teaches mostly 5th-grade students.

To avoid this, our weights also take into account the number of students assigned to a teacher or school who are included in the value-added model.[20] For example, more weight is placed on Teacher A's 4th-grade PSSA math VAM than her 5th-grade PSSA math VAM, even if the precision of the VAMs is 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 (who are included in our models) contribute more heavily to the composite.

In the second step, we placed the assessment-level VAMs on the same scale by normalizing each VAM distribution to have the same standard deviation.[21] Prior value-added studies, including Mathematica analyses using PPS data, have found that the standard deviation of VAM distributions can vary across measures. For example, the standard deviation tends to be slightly larger in math than in reading. It can also vary across grades within a subject. When not normalized, a simple average (for example, an average of a school's VAMs in grade 6 and 7 based on the math PSSA) will implicitly give more weight to the distribution with the larger standard deviation. For example, the VAM score of a top-performing school according to the measure with the larger standard deviation will be farther from the average value, and thus larger than the VAM of a top-performing school according to the other measure. By normalizing the VAM distributions, we ensure that the scores are on the same scale so that the composite weighting methodology determined by district policy can be accurately applied. We use the weights calculated in the first step to average the normalized assessment-level VAM scores and create the composite VAM.

[20] 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) VAMs to determine the weight that assessment receives in the teacher (school) composite VAM.

[21] The shrinkage adjustment described in Chapter III is also included in the composite calculations. Because assessment-level VAMs are reported to teachers, we apply the shrinkage adjustment prior to the normalization step for teacher VAMs. Only composite VAMs are reported to schools, so we apply the shrinkage adjustment after the school composites are calculated.

In the final step, the composites are normalized within the distribution that includes all schools and teachers in the sample and reported as NCEs. The school composite NCE therefore represents each school's performance relative to other schools in Pennsylvania, and the teacher composite NCE represents each teacher's performance relative to other teachers in PPS. In the next chapter, we describe how the effectiveness of other schools in Pennsylvania is measured and is used in the construction of school composite VAMs.

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

Pittsburgh Public Schools seeks VAMs 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-PPS 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 PPS's own data. Most importantly, PPS has data on many student outcomes that are not available statewide, including PSAT results, attendance, and progress in completing core courses. In addition, PPS has more data on students that can be used to improve the predictions of their likely performance. Relying exclusively on statewide data would therefore 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 use a hybrid approach that capitalizes on the breadth of the statewide data and the richness of PPS'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 calculate within-district VAMs to produce fine-grained assessments of how PPS schools performed relative to each other, and we use statewide VAMs to assess where the district as a whole fell in the statewide distribution of performance. This produces a crosswalk or indirect comparison of VAMs that allows us to assess 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. Although all schools in PPS have data on at least one assessment that is given statewide, not all teachers can be assessed with assessments that are available statewide.

Placing PPS schools in the statewide distribution involves two steps. In the first step, we use data on students across Pennsylvania to calculate statewide school VAMs based on PSSA and Keystone Exams. Individual VAMs are calculated and then combined into composites. Second, we assign a statewide NCE to each PPS school based on (1) the distribution of PPS schools in a corresponding statewide composite VAM and (2) the finer-grained VAM rankings produced using the district-specific data. In other words, if the statewide analysis of Keystone data tells us that the top-performing PPS high school in math had a statewide value-added NCE of 90 points, then the school that we identify as top-performing in PPS based on both district and state assessments is assigned to have an NCE score of 90 points. Depending on how well it did on math PSAT compared to the Keystone, that school may or may not be the same one that received a VAM of 90 NCE points based only on math Keystone 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 the PSATs, that are not taken by all students in Pennsylvania. Through the latter goal, we incorporate information on additional measures that cover a broader set of grades and content than could be covered by statewide assessments alone. The two-step process also allows us to use finer-grained background variables available in PPS but not available statewide.

VAMs for PPS include both state and locally administered assessments, but the available state distributions to which these measures can be compared are based on state assessments alone. Our method assumes that the placement of PPS 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 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 calculate statewide VAMs that resemble those described in the preceding chapters as closely as possible given the available data.[22] The statewide VAMs 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 and the PSAT exams are given only to a subset of students outside PPS.
- **Sample includes more students.** The sample size for each statewide VAM is substantially larger than it is when estimating a VAM based on PPS students only. The eligible sample for each individual statewide VAM includes 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 leads to more precise VAMs because we are able to measure the relationships between student characteristics and achievement more accurately.
- **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 cannot include information on gifted participation, course type, prior-year absences, prior-year suspensions, or prior full-year district membership.[23] To limit the potential bias associated with including fewer background characteristics, we add a control for students' own test scores in the same subject from the second prior grade as a third baseline score when available.

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

[23] We exclude the gifted program participation field in the Pennsylvania Information Management System from the state VAM analyses because we are concerned about its validity. There are discrepancies between PPS's gifted participation rate when comparing the state data to PPS data.

- **Less exact dosage measure.** Because data on midyear 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 determine 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 Public School data

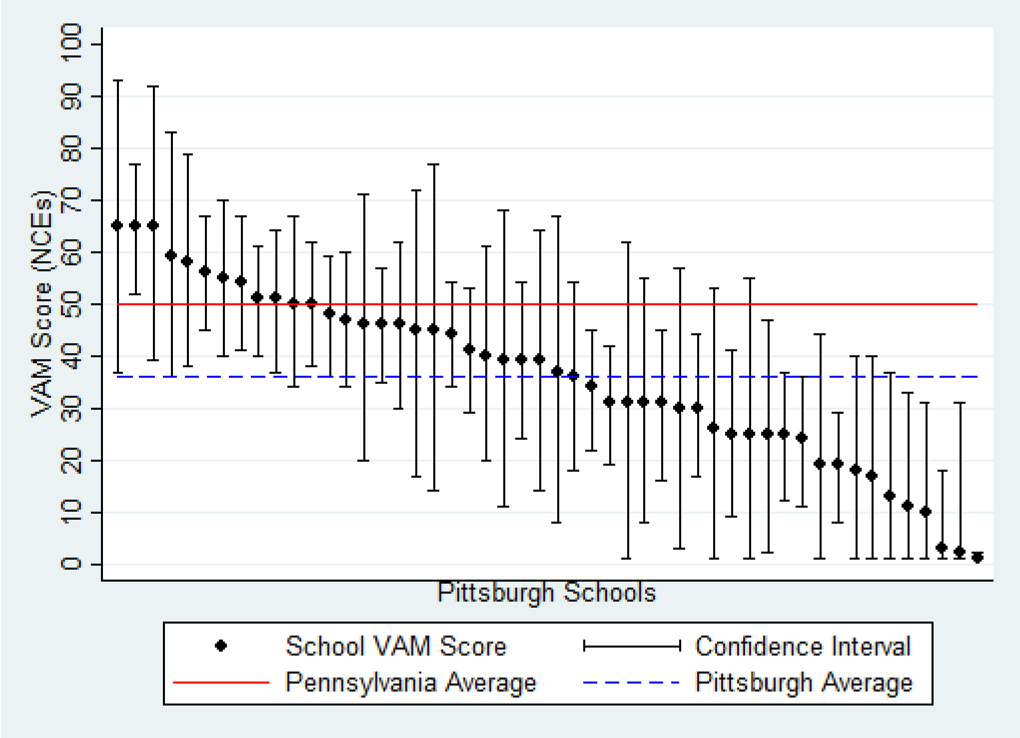
Based on how the distribution of performance in PPS falls relative to the state, our final step is to assign a state value-added NCE to the Pittsburgh VAMs. The distribution of PPS-specific VAMs is adjusted to match the distribution of estimated PPS value added in the statewide analyses. This process attempts to make the most of the available information: statewide VAMs are used to determine the general ranking of PPS schools' performance in the state, and PPS-specific VAMs 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.[24]

All schools, regardless of grade configuration, are 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. Because each VAM is calculated separately by assessment and grade, and as a result of the normalization of VAMs described in Section V.B, a school's place in the statewide distribution is 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 PPS schools (K–5, K–8, 6–8, 6–12, and 9–12) make it difficult to compare schools only to other schools with the same grade configurations when determining the NCE rank. We therefore place 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 VAM. More PPS schools rank below the statewide average in terms of overall value added, though a few exceed the statewide average. The average PPS school received an NCE of 36 (roughly the 25th percentile) in the statewide distribution using data from 2013–14 and 2014–15. The composite VAMs of PPS schools ranged from 1 to 65 NCE points (or the 1st to 76th percentiles), spanning much of the statewide distribution. By comparison, using data from 2011–12 and 2012–13, the average Pittsburgh school received an NCE of 42 (approximately the 35th percentile) (Rotz et al. 2014).

[24] Although school VAMs are adjusted using the statewide data, we use 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 are too narrow and would overstate our true level of precision.

Figure VI.1. Distribution of school composite VAMs in Pennsylvania, 2014–15



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. VAMs 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 PPS. Table VI.1 lists the assessments available statewide, categorized by subject-level composite. PPS-specific composites that include local assessments alongside PSSAs and Keystones are 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 composed of a smaller set of test scores. For example, the 2014–15 PPS elementary math composite included VAMs based on grades 3, 4, and 5 PSSA math scores. This composite was matched to a statewide math composite that includes only PSSA scores for grades 4 and 5. In middle school grades, the state composites include only PSSA VAMs, whereas the district composites include both PSSA VAMs and CBA VAMs. Similarly, district composite VAMs at the high school level incorporated information from Keystone Exams, CBA exams, and the PSAT; state-level composites use only Keystone Exams. Finally, because no social studies exams are available statewide, 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. In future years, PPS VAMs will include fewer district-specific assessments, increasing the overlap between the district-level and statewide composite measures (see Table II.1).

Table VI.1. The composition of statewide composites, 2014–15

	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 PA Core (4, 5) PSSA math (4, 5)	PSSA math PA Core (6, 7, 8) PSSA math (6, 7, 8) Keystone algebra I (8)	Keystone algebra I (9)
English/language arts composite	PSSA ELA PA Core (4, 5) PSSA reading (4, 5) PSSA writing (5)	PSSA ELA PA Core (6, 7, 8) PSSA reading (6, 7, 8) PSSA writing (8)	Keystone literature (10)
Science composite	PSSA science (4)	PSSA science (8)	Keystone biology (9)
Social studies composite	n/a	PSSA ELA PA Core (6, 7, 8) PSSA reading (6, 7, 8) PSSA writing (8)	Keystone literature (10)

Source: 2014–15 Pittsburgh Public Schools value-added models.

Note: The grade level of the majority of students follows each assessment in parentheses. The two-year school VAMs for 2014–2015 included the PSSA assessments aligned to both the legacy and new Pennsylvania Core standards, in addition to the Keystone exams. Starting in 2015–16, they will only include the new PA Core-aligned PSSAs and the Keystone exams.

n/a = not applicable.

Using these mappings, we find that PPS schools tend to perform below the state average. Altogether, 38 percent of Pittsburgh schools have overall composite VAMs statistically different from the average school statewide for the two-year period comprising the 2013–14 and 2014–15 school years: 1 school had above-average statewide performance, and 18 had below-average statewide performance. As shown in Figure VI.1, these schools have confidence intervals that do not include the average value of 50. In comparison, using data from 2011–12 and 2012–13, 30 percent of Pittsburgh schools had overall composite scores statistically different from the average school in the state. Among these, 1 Pittsburgh school had above-average and 14 had below-average statewide performance (Rotz et al. 2014).

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

VII. THE DISTRIBUTION OF TEACHER AND SCHOOL VALUE ADDED IN PITTSBURGH PUBLIC SCHOOLS

In this chapter, we examine the distribution of assessment-level and composite school and teacher VAMs across PPS. This analysis allows us to better understand how the efficacy of teachers and schools varies within the district. 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 2014–15 teacher VAMs are displayed by grade and assessment in Table VII.1. Additional summary results for discontinued assessments are presented in Table A.3 in the appendix. On average across all assessments in 2014–15, we can distinguish 30 percent of teachers from the PPS average using a 95 percent confidence interval. The dispersion of VAMs, measured by their standard deviation, varies by grade, subject, and assessment.[25] At the extremes in Table VII.1, the 90th-percentile teacher raised achievement on the Spanish Multimode Level 2 exam by 0.41 standard deviations compared to the average teacher, whereas the 90th-percentile teacher raised achievement on the 6th-grade PA Core-aligned ELA PSSA by 0.13 standard deviations.

Table VII.1. Teacher VAM results for outcomes, 2014–15

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% level)
PSSA ELA PA Core	4	0.81	1	49	0.20	0.16	0.08	0.29
PSSA Math PA Core	4	0.81	1	44	0.25	0.19	0.08	0.34
PSSA ELA PA Core	5	0.81	1	44	0.20	0.16	0.07	0.30
PSSA Math PA Core	5	0.82	1	42	0.26	0.21	0.08	0.50
PSSA ELA PA Core	6	0.80	1	39	0.13	0.10	0.07	0.13
PSSA Math PA Core	6	0.80	1	35	0.25	0.20	0.08	0.37
PSSA Math PA Core	7	0.75	1	36	0.18	0.14	0.07	0.33
PSSA ELA PA Core	7	0.78	1	40	0.26	0.20	0.08	0.25
Keystone Algebra I	8	0.72	3	25	0.35	0.27	0.12	0.40
PSSA Math PA Core	8	0.62	1	26	0.18	0.14	0.07	0.31
PSSA ELA PA Core	8	0.79	1	34	0.20	0.16	0.07	0.29
Keystone Algebra I	9	0.59	3	39	0.23	0.18	0.09	0.13
Keystone Biology	9	0.72	2	25	0.22	0.17	0.11	0.08
Spanish Multimode Level 2	10	0.44	3	17	0.41	0.32	0.12	0.35

Source: Authors' calculations based on data provided by Pittsburgh Public Schools.

SD = standard deviation.

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

The overall composite VAM for teachers is calculated using the test-based VAMs described by Table VII.1. Overall composite VAMs are available for 575 teachers, of whom 42 percent can be distinguished statistically from typical performance using a 95 percent confidence level. This value is similar to the 37 percent rate reported in Rotz et al. (2014) for the 2012–13 school year.

B. Stability of PSSA VAMs

Starting in the 2014–15 school year, new Math and ELA PSSA exams aligned to PA Core standards replaced the Math, Reading, and Writing PSSAs aligned to legacy standards administered in previous years. As discussed in Section II, because of the difference in content and scale between the new and legacy Math and ELA PSSAs, we calculated separate VAMs for these assessments, which in turn contributed to teachers' composite VAMs. We examined the extent to which composite VAMs for teachers with PSSA VAMs were affected by the introduction of the new PSSAs.

First, as expected, there was no change in average composite scores for teachers with PSSA VAMs. The average VAM for teachers will always be around 50 each year because PPS teacher VAMs are calculated relative to other teachers in the district, and the average teacher performance is set to be equal to 50. The average score for teachers who receive VAM reports is often slightly above 50, because 50 is the average for all teachers in the district over the last three years and it is often the case that less effective teachers leave the district and do not receive reports in the most recent year. Table VII.2 shows the average composite VAM for teachers with at least one grade 3–8 PSSA VAM on their report (excluding Science PSSAs because there was no change to content covered by these exams). The average is close to 51 for each of the last three years.

Table VII.2. Average composite scores for teachers with PSSA VAMs 2012–13 to 2014–15

Year	Average composite VAM
2012–13	51.0
2013–14	51.4
2014–15	51.3

Source: Authors' calculations based on data provided by Pittsburgh Public Schools.

Second, the stability of composite VAMs for teachers with PSSA VAMs was similar over the last two years. One way to study the stability of teacher VAMs is to divide the distribution into quartiles and examine how much movement there is of teachers across performance quartiles. The movement of teachers with at least one PSSA VAM across performance quartiles between 2012–13 and 2013–14 is displayed in Table VII.3. In one extreme, if results were perfectly stable across years, then there would be values of 100 on the diagonals and 0s elsewhere. In the other extreme, if composite VAMs were completely random, there would be values of 25 in each cell of the table.

The actual change in the distribution of teachers falls between these two extremes. The majority of teachers are in the same performance quartile in 2012–13 and 2013–14, and few teachers move by more than one quartile. Table VII.4 shows the change in the distribution between 2013–14 and 2014–15. Overall, the fraction of teachers moving across quartiles in Table VII.4 is similar to Table VII.3. Another way to examine the stability of teacher VAMs is to examine the correlation between composite VAMs across years. The correlation of composite VAMs for teachers with PSSA VAMs between 2012–13 and 2013–14 is 0.828. That same correlation between 2013–14 and 2014–15 is 0.831. In other words, the change in scales and assessments did not reduce the stability of teachers' VAMs.

Table VII.3. Stability of composite scores for teachers with PSSA VAMs: 2012–13 to 2013–14

		2013–14 composite VAM (percentage of teachers)			
		Quartile 1	Quartile 2	Quartile 3	Quartile 4
2012–13 composite VAM (percentage of teachers) N = 353	Quartile 1	65	28	7	0
	Quartile 2	16	49	33	2
	Quartile 3	2	20	54	24
	Quartile 4	1	3	21	76

Source: Authors' calculations based on data provided by Pittsburgh Public Schools.

Table VII.4. Stability of composite scores for teachers with PSSA VAMs: 2013–14 to 2014–15

		2014–15 composite VAM (percentage of teachers)			
		Quartile 1	Quartile 2	Quartile 3	Quartile 4
2013–14 composite VAM (percentage of teachers) N = 315	Quartile 1	71	27	2	0
	Quartile 2	23	52	21	4
	Quartile 3	2	26	52	19
	Quartile 4	0	6	25	69

Source: Authors' calculations based on data provided by Pittsburgh Public Schools.

C. School VAM results

In this section, we first discuss the results of the district-wide school VAMs based on individual assessments and grouped by grade level. Although actual composite VAMs are 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 4–5

The results of the district-wide school VAMs for grades 4 to 5 in 2014–15 are presented in Table VII.5. Summary results for discontinued assessments in the elementary grades are presented in the appendix (see Table A.4). Because of the change in the PSSA assessments, each school-level VAM in math and ELA was based on a single year of data. Across the assessments in Table VII.5 in this grade range, the 90th-percentile school raised achievement by 0.10 to 0.32 standard deviations compared to the average school. For the 2012–13 school year, Rotz et al. (2014) reported differences ranging from 0.14 to 0.34 standard deviations.

Table VII.5. School VAM results for grades 4 to 5, 2014–15

Outcome	Grade	Adj. R-squared	Years of data	Students	Schools	Difference between 90th- and 50th-percentile in z-score units	SD of school effects	Mean standard error	Share of VAMs statistically significant (95% level)
PSSA ELA	4	0.81	1	1,450	34	0.19	0.15	0.10	0.12
PSSA Math Core	4	0.80	1	1,452	34	0.25	0.20	0.11	0.09
PSSA Science	4	0.77	2	2,979	34	0.10	0.08	0.06	0.06
PSSA ELA	5	0.81	1	1,452	34	0.19	0.15	0.11	0.09
PSSA Math Core	5	0.83	1	1,460	34	0.32	0.25	0.12	0.29

Source: Authors’ calculations based on data provided by Pittsburgh Public Schools.

SD = standard deviation.

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.6. Summary results for discontinued assessments in these grades are presented in the appendix (see Table A.5). Many of these models were based on two years of data, though again due to the change in the PSSA exams, those VAMs were calculated using only one year of data. 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 8th-grade Keystone Algebra I test by 0.46 standard deviations compared to the district average, whereas the 90th-percentile school on the 7th-grade PA Core-aligned math PSSA improves results by only 0.15 standard deviations relative to a typical PPS school. These findings are similar to those presented in Rotz et al. (2014) using data from 2011–12 and 2012–13, where the range of gains in grades 6–8 was between 0.10 and 0.62 standard deviations.

Table VII.6. School VAM results for grades 6 to 8, 2014–15

Outcome	Grade	Adj. R-squared	Years of data	Students	Schools	Difference between 90th- and 50th-percentile in z-score units	SD of school effects	Mean standard error	Share of VAMs statistically significant (95% level)
PSSA ELA	6	0.80	1	1,325	22	0.19	0.15	0.08	0.23
PSSA Math Core	6	0.79	1	1,298	22	0.30	0.23	0.10	0.36
PSSA ELA	7	0.78	1	1,316	23	0.29	0.23	0.09	0.26
PSSA Math Core	7	0.75	1	1,212	23	0.15	0.12	0.09	0.04
Keystone Algebra I	8	0.74	2	985	17	0.46	0.36	0.14	0.41
PSSA ELA	8	0.79	1	1,352	23	0.21	0.16	0.09	0.22
PSSA Math Core	8	0.62	1	783	21	0.26	0.20	0.12	0.24
PSSA Science	8	0.79	2	2,889	24	0.18	0.14	0.06	0.42

Source: Authors' calculations based on data provided by Pittsburgh Public Schools.

SD = standard deviation.

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.7. Summary results for discontinued assessments in these grades are presented in the appendix (see Table A.6). Most of the models examining CBAs and Keystones used two years of data, whereas those for the PSAT and CBA Chemistry exam used a single year of data.[26] As in lower grades, the range of school effects varies across subjects. The 90th-percentile PPS high school raises 10th graders' achievement on the PSAT Writing exam by 0.53 standard deviations compared to the average PPS school, but this difference is only 0.07 standard deviations on the 11th-grade PSAT reading exam. These values are similar to those reported in Rotz et al. (2014). In this earlier report, differences between the 90th percentile and average PPS schools ranged from 0.07 to 0.57 standard deviations in 2012–13.

[26] 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 fall 2015 are attributed to the school a student was enrolled in during the 2014–15 academic year. However, these scores are not available in time for inclusion in the model estimation, so only one year of PSAT data (exams taken in fall 2015) was available to use in 2014–15 VAMs.

Table VII.7. School VAM results for grades 9 to 12, 2014–15

Outcome	Grade	Adj. R-squared	Years of data	Students	Schools	Difference between 90th- and 50th-percentile in z-score units	SD of school effects	Mean standard error	Share of VAMs statistically significant (95% level)
Keystone Algebra I	9	0.61	2	1,246	9	0.16	0.12	0.07	0.22
Keystone Biology	9	0.71	2	1,113	8	0.09	0.07	0.08	0.00
Keystone Literature	10	0.68	2	1,285	9	0.11	0.09	0.06	0.00
PSAT Math	10	0.71	1	960	9	0.14	0.11	0.05	0.56
PSAT Reading	10	0.67	1	944	9	0.14	0.11	0.06	0.33
PSAT Writing	10	0.64	1	944	9	0.53	0.41	0.06	0.67
PSAT Math	11	0.85	1	1,049	9	0.09	0.07	0.05	0.11
PSAT Reading	11	0.84	1	1,012	9	0.07	0.06	0.05	0.00
PSAT Writing	11	0.76	1	1,012	9	0.43	0.34	0.07	0.78

Source: Authors' calculations based on data provided by Pittsburgh Public Schools.

SD = standard deviation.

4. Composite school VAM results

After estimating all the assessment-level school VAMs, we combine them into four subject-level composites, an overall test-based composite, and two non-test composites (see Chapter V). We then map the within-district performance of PPS 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.10. We report these statistics for the attendance and core course pass rate VAMs as well, though for the non-test VAMs, comparisons are 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.

In 2014–15, 40 percent of PPS schools could be statistically distinguished from the average Pennsylvania school based on math scores, 10 percent based on reading scores, 67 percent based on science scores, and 57 percent based on social studies scores. Using the test-based composite VAMs, 38 percent of PPS schools are significantly better or worse than the average school in the state. We can distinguish 38 percent of PPS schools based on attendance and 56 percent of schools based on core course pass rates from the average school in Pittsburgh.

Table VII.8. 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	23	0.61	0.04	0.17	n/a	0.30	0.13	n/a
K-8	11	0.09	0.09	0.55	0.55	0.27	0.27	n/a
6-8	7	0.29	0.29	1.00	0.25	0.43	0.86	n/a
6-12	5	0.40	0.20	0.60	0.67	0.60	1.00	0.40
9-12	4	0.25	0.00	0.50	0.50	0.75	0.50	0.75

Source: Authors' calculations based on data provided by Pittsburgh Public Schools.

n/a = not applicable.

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 understand how and why achievement varies across PPS students. This understanding can further inform decisions made by the PFT, PPS, and other key stakeholders.

The 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 (nongifted students), holding all else equal. For example, on average across the VAMs, we find that gifted students typically score 0.11 standard deviations higher than nongifted 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, including their prior test scores), the gifted student is predicted to score 0.11 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.[27] We summarize the average relationship between the control variable and outcome 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 statistically 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).

Many of the control variables are strongly correlated, which can result in statistically insignificant coefficients because there is little independent variation to estimate the relationship between a control variable and the outcome test score when holding the other control variables constant. For example, students who were frequently suspended during the prior year may also have many prior-year absences, so it may be difficult to precisely estimate the independent relationship between these two variables. This issue, known as multicollinearity, does not affect the predictive power of the model but can result in noisy estimates of some of the control variable coefficients. Furthermore, it is possible for a coefficient to be significantly positive and negative in some of the models, even if on average there is no relationship between that control

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

variable and test scores, as we are examining results across several different models and a small number of statistically significant results can occur due to chance.

The results in Table VIII.1 indicate relationships between student characteristics and outcomes that are consistent with previous research. 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, their relative performance varies by subject. Across the 2014–15 PSSA exams, for example, male students performed significantly worse than female students in all grades of the ELA exams. On the other hand, they performed about the same as female students in the math exams and significantly better in the science exams. Students from low-income families, as measured by free or reduced-price lunch status, also perform significantly worse than other students in 25 percent of the value-added models.

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. English-language learners who gain English skills between the baseline and outcome tests are likely to improve their performance even more than otherwise equivalent students, as they learn English and catch up to their peers. Likewise, students who repeated the assessments have higher test scores than other students in about one third of the value-added models. Students who missed many days 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 during the current year; that is, students who were absent for many days in the past school year may be more likely to miss school in the current year, which can bring down current-year test scores.

Table VIII.1. Relationship between student 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)**
Race (African American is reference category)			
White	0.04	0.29	0.02
Hispanic	0.06	0.10	0.00
Asian	0.19	0.43	0.00
Other race	0.05	0.16	0.00
Ever applied to magnet school	0.02	0.14	0.00
Male	-0.04	0.21	0.41
Lunch program	-0.05	0.00	0.25
English-language learner	0.10	0.22	0.02
Gifted	0.11	0.40	0.03
Moved schools in current year	-0.09	0.02	0.35
Past year proportion absent	-0.21	0.02	0.21

	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)
Past year proportion suspended	-0.56	0.03	0.05
In PPS all of past year	-0.04	0.02	0.13
Age in years	-0.04	0.00	0.25
Specific learning disability	-0.02	0.08	0.15
Speech or language impairment	-0.04	0.06	0.10
Emotional disturbance	-0.04	0.05	0.05
Intellectual disability	-0.02	0.12	0.10
Autism	0.09	0.21	0.05
Physical/sensory impairment	0.02	0.10	0.02
Other impairment	-0.04	0.02	0.13
Repeated assessment	0.11	0.38	0.14
Pittsburgh Scholars Program	-0.05	0.00	0.25
Advanced Placement Program	-0.05	0.00	0.00
Center for Advanced Studies Student	0.12	0.00	0.00

Source: Authors' calculations based on data provided by Pittsburgh Public Schools.

Note: Estimated relationships from all teacher-level, test-based, value-added models from 2014–15 including the relevant characteristic.

In addition to student-level characteristics, we include classroom-average characteristics in the VAMs. Table VIII.2 summarizes these estimated relationships. Class size has the most consistent relationship with student test scores; the coefficient is statistically significant and negative in 40 percent of value-added models. An increase in class size of one student is associated with a decrease in student test scores of about 1 percent of one standard deviation. Generally speaking, fewer of the coefficients on classroom characteristics are consistently significant and of the same sign. For example, the relationship between student test scores and the fraction of gifted students in the class is significant and positive in 12 percent of value-added models but significant and negative in 9 percent. Overall, student-level characteristics tend to have more consistent relationships with predicted test scores than average classroom 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	0.05	0.09	0.02
Share Hispanic	0.19	0.19	0.05
Share Asian	0.30	0.21	0.07
Share other race	−0.03	0.02	0.05
Share male	0.00	0.07	0.12
Share in lunch program	−0.09	0.02	0.26
Share English language learners	−0.08	0.05	0.09
Share gifted	0.15	0.12	0.09
Share with any disability	−0.08	0.05	0.14
Average past year proportion absent	0.36	0.00	0.05
Average past year proportion suspended	−0.37	0.12	0.12
Share in PPS all of past year	−0.05	0.05	0.16
Class size	−0.01	0.02	0.40
Past year math score	0.00	0.19	0.14
Past year reading score	0.05	0.16	0.07

Source: Authors' calculations based on data provided by Pittsburgh Public Schools.

Notes: Estimated relationships from all teacher-level, test-based value-added models including the relevant characteristic. The past year scores used in these models are summarized in Table II.3.

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 composite VAMs 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 VAM 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 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 include only outcomes available statewide. Because STAR requires that only statewide assessments be used, these VAMs are calculated using only PSSA scores from grades 4 to 8 and Keystone Algebra I, Biology, and Literature scores. We calculate assessment-level statewide VAMs and develop for each school grade range (4–5, 6–8, and 9–12) a single composite measure including all the relevant state assessments. Schools with grade configurations of K–8 or 6–12 receive composite VAMs that include all STAR outcomes from those grades. We then use the composite VAMs to determine which schools place in the top 15 (or 25) percent of the statewide distribution. To simplify the process for displaying 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 to identify STAR schools in each grade range for 2014–15. STAR awards were determined based on the overall composite VAMs. Using data from 2013–14 and 2014–15, four Pittsburgh schools were found to be in the top 25 percent of the statewide distribution. No schools placed in the top 15 percent. Table IX.2 displays the weights each subject-level composite received when constructing the overall STAR composite VAM. These weights were determined by the district.

Table IX.1. Assessments used to determine the STAR award system by grade range, 2013–15 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 math PA core (4, 5)*	PSSA math PA Core (6, 7, 8)*	
	PSSA reading (4, 5)*	PSSA reading (6, 7, 8)*	Keystone literature (10)
	PSSA ELA PA core (4, 5)*	PSSA ELA PA core (6, 7, 8)*	
	PSSA writing (5)*	PSSA writing (8)*	
	PSSA science (4)	PSSA science (8)	Keystone biology (9)

Source: 2014–15 Pittsburgh Public Schools value-added models.

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

*Based on one year of data

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	0.43	0.35
Reading	0.46	0.40	0.44
Science	0.10	0.13	0.21
Writing	0.07	0.04	n/a
Total	**1.00**	**1.00**	**1.00**

Source: 2014–15 Pittsburgh Public Schools value-added models.

n/a = not applicable.

APPENDIX A

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

Table A.1 lists the assessments used in 2014–15 VAMs that PPS plans to discontinue beginning in 2015–16. The remaining tables in the appendix provide additional information and results for these discontinued assessments.

Most of the exams listed in Tables A.1 were used in both 2014–15 teacher and school VAMs. One exception is the 11th-grade CBA ELA exam. We were unable to calculate teacher VAMs for this assessment due to an issue with the roster verification process in 2014–15 that prevented teachers from verifying their student enrollment and instructional responsibility for this assessment. We excluded the assessment from teacher VAMs entirely rather than relying on the first two years in the three-year window because using just two years of data did not provide precise enough results to be useful for teacher VAMs.

Due to changes in assessments and curricular standards over the periods spanned by school and teacher VAMs, some of the assessments listed in Table A.1 did not contribute data from the 2014–15 school year. Because PPS ceased administering the grade 8 physical science CBA and grade 2 TerraNova exams, these assessments only contributed one year of data to the three-year teacher VAMs and were not included in the two-year school VAMs. The grade 8 physical science CBA was last administered during the 2012–13 school year, and the grade 2 TerraNova exams were last administered during the 2013–14 school year. However, PPS made the policy decision not to use TerraNova scores from the last year the exams were administered as outcomes for accountability purposes (the exams were still used as baseline controls for the 2014–15 grade 3 PSSA value-added models).

In other cases, the alignment between the assessments and course curricula changed. PPS determined that the content covered by the grade 4 science PSSA was taught across multiple grades beginning in 2013–14, and thus this assessment was no longer used in teacher VAMs starting that year. PPS similarly determined that the grade 10 chemistry CBA did not directly align with the curriculum taught in 2014–15, and thus should not be used for either school or teacher VAMs that year.

Table A.1. Test scores last used in 2014–15 VAMs, by subject and grade

Test	Grade										
	2	3	4	5	6	7	8	9	10	11	12
State assessments											
PSSA ELA PA Core		S,T									
PSSA Math PA Core		S,T									
PSSA Reading		S*,T*	S*,T*	S*,T*	S*,T*	S*,T*	S*,T*				
PSSA Writing				S*,T*			S*,T*				
PSSA Math		S*,T*	S*,T*	S*,T*	S*,T*	S*,T*	S*,T*				
PSSA Science			T*								
Standardized tests not given across state											
TerraNova Math	T*										
TerraNova Reading	T*										
Locally developed assessments											
CBA Math					S,T	S,T	S,T				
CBA Algebra I							S,T	S,T			
CBA Geometry									S,T		
CBA Algebra II										S,T	
CBA Reading					S,T	S,T	S,T				
CBA ELA								S,T	S,T	S	S,T
CBA African American Lit											S,T
CBA Earth Science					S,T						
CBA Life Science							S,T				
CBA Biology								S,T			
CBA Chemistry									S*,T*		
CBA Physics										S,T	
CBA Physical Science						S,T	T*				
CBA Civics								S,T			
CBA World Geography					S,T						
CBA World History						S,T			S,T		
CBA US History							S,T			S,T	

Source: 2014–15 Pittsburgh Public Schools value-added models.

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. The Reading, Writing, and Math PSSAs not noted as “PA Core” correspond to those aligned to legacy standards, administered prior to 2014–15. When students take both a midterm and final CBA, we combine both scores by placing one-third weight on the midterm score and two-thirds weight on the final score. If the midterm score is missing, the final score is used as the outcome. If the final score is missing, the student is excluded from that value-added model.

*Indicates that the assessment did not contribute data from the 2014–15 school year to that VAM.

Table A.2. Discontinued assessment outcomes and baseline test scores

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 PA Core Grade 3	TerraNova Math Grade 2	TerraNova Reading Grade 2	
PSSA ELA PA Core 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 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 World Geography 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 Physical Science Grade 7	CBA World Geography Grade 6***	PSSA Math Grade 6	PSSA Reading Grade 6
CBA World History Grade 7	CBA Earth Science Grade 6	PSSA Math Grade 6	PSSA Reading Grade 6
CBA Algebra I Grade 8	PSSA Math Grade 7	PSSA Reading Grade 7	
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 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	
CBA Physical Science Grade 8**	CBA Life Science Grade 7	PSSA Math Grade 7	PSSA Reading Grade 7
CBA Life Science Grade 8	CBA Physical Science Grade 7	PSSA Math Grade 7	PSSA Reading Grade 7
CBA US History Grade 8	CBA World History Grade 7***	PSSA Reading Grade 7	PSSA Math Grade 7
CBA 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
CBA Geometry Grade 10	CBA Algebra I 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
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
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
CBA ELA IV/African American Lit Grade 12	CBA ELA III Grade 11	PSSA Reading Grade 8	PSSA Writing Grade 8

Source: 2014–15 Pittsburgh Public Schools value-added models.

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

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

***This exam was used as an additional baseline control variable in 2013–14 and 2014–15. It was not available as a baseline in 2012–13.

Table A.3. Teacher VAM results for discontinued outcomes, 2014–15

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% level)
TerraNova Reading	2	0.69	1	67	0.26	0.20	0.11	0.19
TerraNova Math	2	0.75	1	64	0.32	0.25	0.10	0.33
PSSA ELA PA Core	3	0.73	1	55	0.17	0.13	0.09	0.11
PSSA Math PA Core	3	0.70	1	45	0.21	0.16	0.09	0.20
PSSA Math	3	0.73	2	66	0.24	0.18	0.09	0.21
PSSA Reading	3	0.70	2	80	0.19	0.15	0.09	0.14
PSSA Reading	4	0.79	2	71	0.19	0.15	0.09	0.18
PSSA Science	4	0.80	1	40	0.17	0.13	0.07	0.28
PSSA Math	4	0.82	2	65	0.29	0.22	0.09	0.40
PSSA Writing	5	0.57	2	64	0.25	0.19	0.11	0.19
PSSA Reading	5	0.78	2	65	0.13	0.10	0.07	0.03
PSSA Math	5	0.86	2	55	0.24	0.19	0.07	0.40
PSSA Math	6	0.85	1	42	0.19	0.15	0.07	0.29
CBA Reading	6	0.66	3	58	0.21	0.16	0.09	0.22
CBA Math	6	0.69	3	61	0.38	0.29	0.09	0.49
CBA Social Studies	6	0.65	3	43	0.37	0.29	0.12	0.42
PSSA Reading	6	0.76	2	75	0.22	0.17	0.09	0.24
CBA Earth Science	6	0.67	3	58	0.48	0.37	0.09	0.71
PSSA Reading	7	0.78	2	72	0.14	0.11	0.07	0.01
CBA Reading	7	0.60	3	66	0.15	0.12	0.08	0.09
CBA Physical Science	7	0.67	3	45	0.49	0.39	0.08	0.64
CBA Social Studies	7	0.64	3	31	0.40	0.31	0.10	0.61
PSSA Math	7	0.84	2	52	0.17	0.14	0.07	0.19
CBA Math	7	0.66	3	56	0.37	0.29	0.09	0.59
CBA Algebra I	8	0.58	3	23	0.31	0.24	0.13	0.22
PSSA Writing	8	0.57	2	48	0.31	0.24	0.10	0.40
CBA Life Science	8	0.71	2	29	0.54	0.42	0.09	0.69
CBA Reading	8	0.62	3	55	0.28	0.22	0.09	0.33
CBA Math	8	0.45	3	43	0.43	0.34	0.13	0.42
CBA Physical Science	8	0.64	1	21	0.35	0.28	0.08	0.67
CBA US History	8	0.69	3	29	0.30	0.23	0.08	0.41
PSSA Reading	8	0.78	2	47	0.17	0.14	0.07	0.19
PSSA Math	8	0.73	2	40	0.22	0.17	0.09	0.23
CBA Algebra I	9	0.41	3	31	0.32	0.25	0.10	0.29
CBA ELA I	9	0.50	3	34	0.24	0.19	0.10	0.32
CBA Civics	9	0.68	3	31	0.40	0.32	0.10	0.58
CBA Biology	9	0.52	3	24	0.35	0.27	0.10	0.63
CBA Geometry	10	0.48	3	34	0.24	0.19	0.10	0.18
CBA ELA II	10	0.53	3	31	0.46	0.36	0.12	0.42
CBA Civics	10	0.61	3	29	0.47	0.37	0.11	0.41
CBA Biology	10	0.59	2	20	0.36	0.28	0.12	0.25
CBA Physics	11	0.53	3	17	0.53	0.41	0.12	0.47
CBA US History	11	0.56	3	26	0.51	0.40	0.15	0.46
CBA Algebra II	11	0.47	3	36	0.51	0.39	0.12	0.53
CBA ELA IV/African American Lit	12	0.48	3	26	0.36	0.28	0.12	0.35

Source: Authors' calculations based on data provided by Pittsburgh Public Schools. SD = standard deviation.

Table A.4. School VAM results for grades 3 to 5 discontinued outcomes, 2014–15

Outcome	Grade	Adj. R-squared	Years of data	Students	Schools	Difference between 90th- and 50th-percentile in z-score units	SD of school effects	Mean standard error	Share of VAMs statistically significant (95% level)
PSSA Math PA Core	3	0.71	1	1,514	34	0.25	0.20	0.11	0.18
PSSA ELA	3	0.74	1	1,511	34	0.23	0.18	0.11	0.18
PSSA Math	3	0.73	1	1,477	33	0.27	0.21	0.11	0.18
PSSA Reading	3	0.68	1	1,486	33	0.17	0.14	0.10	0.03
PSSA Math	4	0.80	1	1,535	34	0.32	0.25	0.11	0.41
PSSA Reading	4	0.78	1	1,575	34	0.35	0.27	0.10	0.38
PSSA Math	5	0.88	1	1,484	34	0.40	0.31	0.10	0.53
PSSA Reading	5	0.81	1	1,502	34	0.16	0.12	0.09	0.03
PSSA Writing	5	0.62	1	1,429	33	0.29	0.23	0.13	0.12

Source: Authors' calculations based on data provided by Pittsburgh Public Schools.

SD = standard deviation.

Table A.5. School VAM results for grades 6 to 8 discontinued outcomes, 2014–15

Outcome	Grade	Adj. R-squared	Years of data	Students	Schools	Difference between 90th- and 50th-percentile schools in z-score units	SD of school effects	Mean standard error	Share of VAMs statistically significant (95% level)
CBA Earth Science	6	0.66	2	2,474	22	0.60	0.47	0.08	0.86
CBA Math	6	0.68	2	2,318	22	0.38	0.30	0.09	0.68
CBA Reading	6	0.67	2	2,078	21	0.25	0.19	0.08	0.29
CBA Social Studies	6	0.63	2	1,802	18	0.40	0.32	0.10	0.56
PSSA Reading	6	0.78	1	1,460	24	0.22	0.17	0.09	0.33
CBA Math	7	0.62	2	2,135	22	0.29	0.23	0.08	0.45
CBA Physical Science	7	0.63	2	2,308	22	0.47	0.37	0.08	0.59
CBA Reading	7	0.59	2	2,108	22	0.11	0.08	0.07	0.00
CBA Social Studies	7	0.70	2	1,569	18	0.51	0.40	0.09	0.72
PSSA Math	7	0.85	1	1,319	24	0.14	0.11	0.07	0.00
PSSA Reading	7	0.79	1	1,435	24	0.10	0.07	0.06	0.00
CBA Algebra I	8	0.64	2	875	16	0.52	0.40	0.17	0.44
CBA Life Science	8	0.68	2	2,253	21	0.49	0.38	0.08	0.62
CBA Math	8	0.47	2	1,361	21	0.49	0.38	0.13	0.62
CBA Reading	8	0.63	2	2,328	22	0.24	0.19	0.08	0.50
CBA US History	8	0.71	2	1,502	16	0.34	0.26	0.09	0.44
PSSA Math	8	0.70	1	848	20	0.24	0.18	0.13	0.15
PSSA Reading	8	0.79	1	1,480	22	0.20	0.16	0.09	0.27
PSSA Writing	8	0.59	1	1,471	22	0.40	0.31	0.12	0.50

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

SD = standard deviation.

Table A.6. School VAM results for grades 9 to 12 discontinued outcomes, 2014–15

Outcome	Grade	Adj. R-squared	Years of data	Students	Schools	Difference between 90th- and 50th-percentile schools in z-score units	SD of school effects	Mean standard error	Share of VAMs statistically significant (95% level)
CBA Biology	9	0.50	2	1,190	6	0.22	0.17	0.07	0.67
CBA ELA 1	9	0.49	2	1,132	8	0.38	0.29	0.11	0.50
CBA Algebra I	9	0.40	2	989	6	0.27	0.21	0.07	0.33
CBA Civics	9	0.64	2	1,094	7	0.37	0.29	0.08	0.57
CBA ELA II	10	0.47	2	1,129	7	0.50	0.39	0.10	0.57
CBA World History	10	0.57	2	957	6	0.25	0.20	0.09	0.00
CBA Chemistry	10	0.60	1	551	6	0.84	0.66	0.14	0.67
CBA Geometry	10	0.49	2	1,145	8	0.21	0.16	0.09	0.38
CBA Physics	11	0.47	2	763	6	0.39	0.30	0.12	0.50
CBA Algebra II	11	0.34	2	964	7	0.28	0.22	0.11	0.43
CBA US History	11	0.52	2	768	6	0.56	0.43	0.13	0.50
CBA ELA III	11	0.46	2	880	7	0.65	0.51	0.12	0.57
CBA ELA IV/African American Lit	12	0.39	2	840	7	0.26	0.21	0.11	0.14

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

SD = standard deviation.

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

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.

Bacher-Hicks, A., T. J. Kane, and D. O. Staiger. "Validating Teacher Effect Estimates Using Changes in Teacher Assignments in Los Angeles." NBER Working Paper No. 20657. Cambridge, MA: National Bureau of Economic Research, November 2014.

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." *American Economic Review*, vol. 104, no. 9, 2014b, pp. 2593–2632.

Chetty, R., J. N. Friedman, and J. E. Rockoff. "Measuring the Impacts of Teachers II: Estimating Bias in Teacher Value-Added Estimates." *American Economic Review*, vol. 104, no. 9, 2014a, pp. 2633–2679.

College Board. "PSAT/NMSQT Understanding Scores 2015." 2016. Available at https://collegereadiness.collegeboard.org/pdf/2015-psat-nmsqt-understanding-scores.pdf.

Goldhaber, Dan, and Duncan Chaplin. "Assessing the 'Rothstein Falsification Test': Does It Really Show Teacher Value-Added Models Are Biased?" *Journal of Research on Educational Effectiveness*, vol. 8, no. 1, 2015, pp. 8–34.

Bloom, H. S., C. J. Hill, A. R. Black, and M. W. Lipsey. "Performance Trajectories and Performance Gaps as Achievement Effect-Size Benchmarks for Educational Interventions." *Journal of Research on Educational Effectiveness*, vol. 1, no. 4, 2008, pp. 289–329.

Guarino, C., M. D. Reckase, and J. M. Woolridge, "Can Value-Added Measures of Teacher Performance Be Trusted?" *Education Finance and Policy*, vol. 10, no. 1, 2014, pp. 117–156.

Hock, H., and E. Isenberg. "Methods for Accounting for Co-Teaching in Value-Added Models." Mathematica Policy Research Working Paper 6. 2012. Available at https://www.mathematica-mpr.com/-/media/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 http://dcps.dc.gov/sites/default/files/dc/sites/dcps/publication/attachments/Measuring%20Value%20Added%20in%20DC%202011-2012.pdf.

Isenberg, Eric, and Elias Walsh. "Accounting for Co-Teaching: A Guide for Policymakers and Developers of Value-Added Models." *Journal of Research on Educational Effectiveness*, vol. 8, no. 1, 2015, pp. 112–119.

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, Cambridge, MA: National Bureau of Economic Research, 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.

Lipscomb S., B. Gill, K. Booker, and M. Johnson. "Assessing Teacher Effectiveness in Pittsburgh Public Schools: Value-Added Modeling and Results" Report to the Pittsburgh Public Schools. Cambridge, MA: Mathematica Policy Research. 2010.

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, Alexandra, and Eric Isenberg. "How Do Test Scores at the Floor and Ceiling Affect Value-Added Estimates?" Working paper 33. Washington, DC: Mathematica Policy Research, July 2014. Available at https://www.mathematica-mpr.com/our-publications-and-findings/publications/how-do-test-scores-at-the-floor-and-ceiling-affect-value-added-estimates.

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.

Rotz, Dana, Matthew Johnson, and Brian Gill. “Value-Added Models for the Pittsburgh Public Schools, 2012–13 School Year.” Report to the Pittsburgh Public Schools. Cambridge, MA: Mathematica Policy Research, 2014. Available at http://www.mathematica-mpr.com/~/media/publications/pdfs/education/vam_pittsburgh.pdf.

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.

www.mathematica-mpr.com

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

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.