Value-Added Models for the Pittsburgh Public Schools

February 2012

Matthew Johnson
Stephen Lipscomb
Brian Gill
Kevin Booker
Julie Bruch

MATHEMATICA
Policy Research

All statistics are calculated by Mathematica unless stated otherwise

Contract Number:
0E9347

Mathematica Reference Number:
06723.300

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

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

Value-Added Models for the Pittsburgh Public Schools

February 2012

Matthew Johnson
Stephen Lipscomb
Brian Gill
Kevin Booker
Julie Bruch

MATHEMATICA
Policy Research

All statistics are calculated by Mathematica unless stated otherwise

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 two years that this work has been underway. Although we cannot name everyone, we want to especially thank Mary Wolfson, Marni Pastor, Paulette Poncelet, Grace Tan, Sam Franklin, Veronica Amundson, Jeannine French, Eddy Jones, and Bill Hileman for their support for the work. Angela Minnici and Rob Weil from the American Federation of Teachers provided useful input early in our 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 Duncan Chaplin, Hanley Chiang, John Deke, and Eric Isenberg for their technical suggestions and Rosie Malsberger and Amanda Hakanson for organizing the data to set up the model estimation. Ryan Sowers ably led the process for producing individual value-added reports for teachers and schools, assisted by John McCauley and Lei Rao. And Eileen Curley formatted the report to make it ready for public release.

GLOSSARY

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

CONTENTS

I. INTRODUCTION 1

II. STUDENT OUTCOMES AND BACKGROUND VARIABLES USED IN PITTSBURGH'S VALUE-ADDED MODELS 5

A. Test Outcomes and Baselines 5

B. Non-Test Outcomes and Baselines 8

C. Student and Peer Characteristics, Class Size, Course Type, and School Choice 10

III. TECHNICAL DETAILS OF PPS VALUE-ADDED MODELS 13

A. Detailed Value-Added Model Description 13

B. Standardization 14

C. Teacher and School Dosage 14

D. Shrinkage 15

E. Technical Details for VAMs for Non-Test Outcomes 15

IV. METHODOLOGICAL LIMITATIONS 17

A. Non-Random Assignment of Students 17

B. Distinguishing Between School and Teacher Effects 17

C. School VAMs Do Not Use "Pre-Treatment" Baselines 18

D. Absence of VAM Estimates for Grades K-3 18

E. Missing and Omitted Data 19

V. COMPOSITE VALUE-ADDED MEASURES 21

A. Composition of Composites 21

B. Construction of Composite Estimates: Precision Weighting 22

VI. SUMMARIZING PITTSBURGH SCHOOL PERFORMANCE IN THE CONTEXT OF A STATEWIDE DISTRIBUTION 25

A. Statewide Teacher and School VAMs 26

B. Assigning a State Value-Added Percentile to Results Based on Pittsburgh Data 27

C. Comparison to PVAAS....29

VII. THE DISTRIBUTION OF TEACHER AND SCHOOL VALUE-ADDED IN PPS....31

A. Teacher VAM Results....31

B. School VAM Results....33

VIII. APPLICATIONS TO REWARDS AND RECOGNITION OPPORTUNITIES....39

A. Promise-Readiness Corps Details....39

B. Students and Teachers Achieving Results (STAR) Details....41

REFERENCES....43

APPENDIX TABLES....A-1

TABLES

II.1	Available Test Scores by Subject and Grade	5
II.2	Assessment Outcomes and Baseline Test Scores Used in PPS VAMs, 2010-11	7
II.3	Non-Assessment Outcomes and Baseline Measures Used in School VAMs, 2010-11	8
II.4	Variables for Student Background Characteristics in Pittsburgh Teacher and School VAMs, 2010-11	11
V.1	The Composition of Subject Composites for Pittsburgh School VAMs, 2010-11	21
VI.1	The Composition of Statewide Composites, 2009-11	29
VII.1	Teacher VAM Results, by Outcome 2008-11	32
VII.2	School VAM Results for Grades 4 to 5, by Outcome	34
VII.3	School VAM Results for Grades 6 to 8, by Outcome	35
VII.4	School VAM Results for Grades 9 to 12, by Outcome	36
VII.5	School VAM Results, Non-test Outcomes	37
VII.6	Test-Based Composite School VAM Results	37
VIII.1	Assessments Used to Determine the STAR Award System by Grade Range, 2011-12	41
A.1	Precision of Teacher VAM Results for Grades 6 to 8, by Years of Teaching Included	A-1
A.2	Correlation of Mathematica and PVAAS VAM Estimates, Grades 4 to 8	A-1

FIGURES

I.1 Prediction Based on Last Year's Score 2

I.2 Prediction Based on Last Year's Score 2

I.3 Prediction Based on Last Year's Score + Gifted + IEP..... 3

VI.1 Distribution of Composite School VAM Estimates in Pennsylvania, 2009-11 28

VI.2 State Percentile Composite VAM Ranking of Pittsburgh Schools, 2009-11 28

VII.1 Teacher VAM Results for Math and Reading PSSAs Expressed in Fractions of a Year of Learning: Difference between Median Teacher and 90th-percentile Teacher in Pittsburgh 33

I. INTRODUCTION

At the request of Pittsburgh Public Schools (PPS) and the Pittsburgh Federation of Teachers (PFT), Mathematica has developed value-added models (VAMs) that aim to estimate the contributions of individual teachers, teams of teachers, and schools to the achievement growth of their students. Our work in estimating value-added in Pittsburgh supports the larger, joint efforts of PPS and the PFT to "empower effective teachers" through evaluation, professional development, and compensation. Pittsburgh's VAMs use not only state assessments but also course-specific assessments, student attendance, and course completion rates, thereby aiming to produce estimates of the contributions of teachers and schools that are fair, valid, reliable, and robust. The findings in this report suggest that the VAM estimates provide meaningful information about teacher and school performance in Pittsburgh. The VAM results for individual schools and teachers have been reported to them privately.

A VAM provides a better indication of effectiveness than average score levels or the rate of student proficiency because it examines the trajectory of achievement for students from a baseline and accounts for other factors that affect student achievement and are outside the control of teachers or schools (Meyer 1997). The process of estimating a teacher or school value-added model can be conceptualized as occurring in two steps. In the first step the VAM makes a prediction about an outcome of interest, typically a student's assessment score in a subject, based on factors including students' own achievement histories and other characteristics of students and their peers. Each student's own prior achievement is the most important element in the prediction. These predictions represent what we would expect the students to achieve if they were served by the average teacher or school. In the second step, the VAM compares students' actual outcomes to their predicted outcomes. The VAM score 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 achievement histories and background characteristics if taught by the "average" teacher or school?*

The predicted achievement level for each student is the best estimate of how that student will do, given everything we know about the student. The predictions we generate are based on data from the current year as well as the past year—we cannot actually predict an outcome in advance, because we need to know how well similar students perform in the current year in order to predict an outcome for any particular student. Figures I.1, I.2, and I.3 provide a simplified graphical illustration of how these predictions work.

Figure I.1 draws a simple prediction line in which a student's 2011 test score is predicted based only on the student's 2010 score in the same subject. The prediction line is derived using the student test score data. Each pair of scores (2010 and 2011) for an individual student is represented by a diamond on the chart. Assume for the sake of simplicity that the diamonds represent all students across the entire district (or state). The red diamonds represent gifted students, and the yellow diamonds represent students with Individualized Education Plans (IEPs), i.e., special education students. If we were predicting the 2011 score of another student, knowing only her 2010 score, we would select the 2011 score that falls on the diagonal line at the point corresponding to the student's 2010 score.

Figure I.1. Prediction Based on Last Year's Score

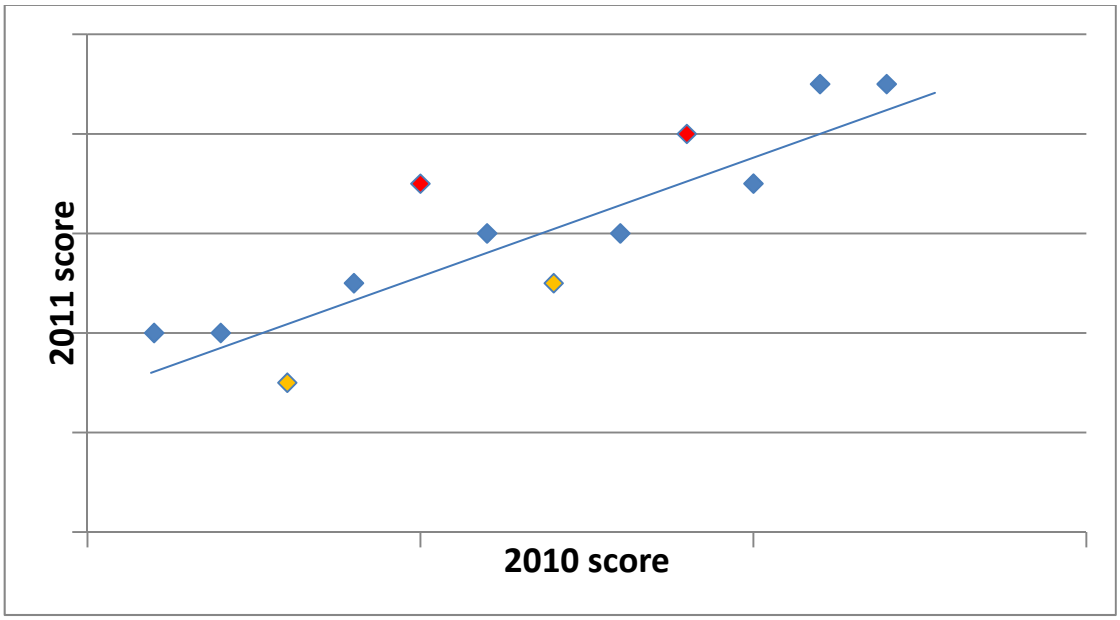


Figure I.2 adds an additional piece of information to last year's score. We know that the two students represented by red diamonds are gifted students, and both of those students have scores above the line. This suggests that gifted students, on average, are predicted to do slightly better in 2011 than non-gifted students who had the same 2010 scores. If we were predicting the 2011 score of another gifted student, we would adjust our prediction upward from the line by an amount that is approximately the average height of the distance between the red diamonds and the blue line. This adjusted predicted line for gifted students is represented in red.

Figure I.2. Prediction Based on Last Year's Score

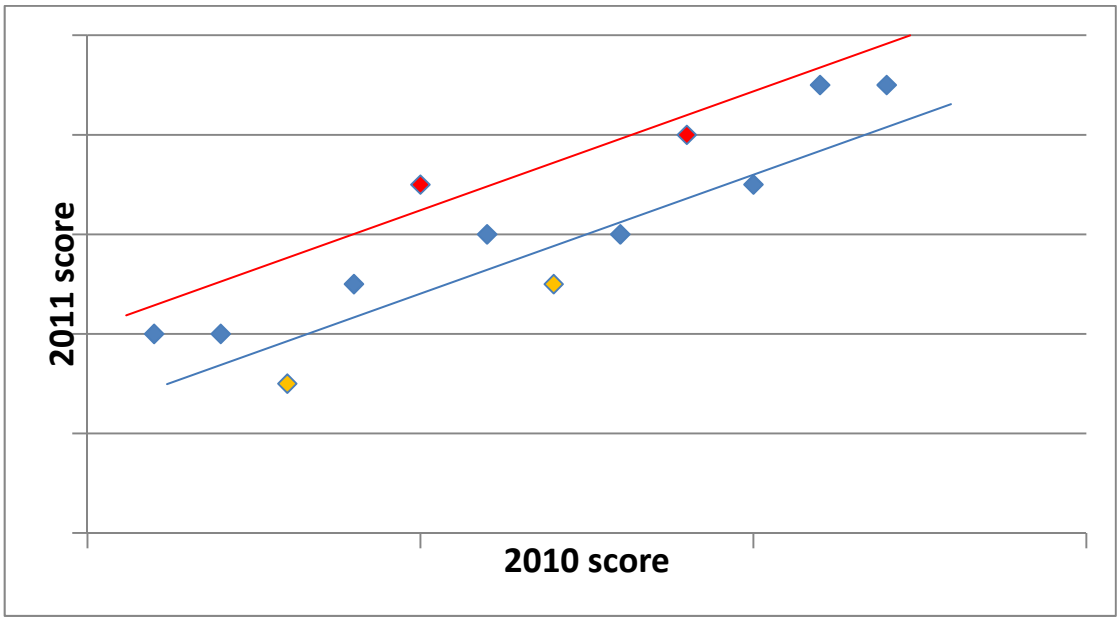
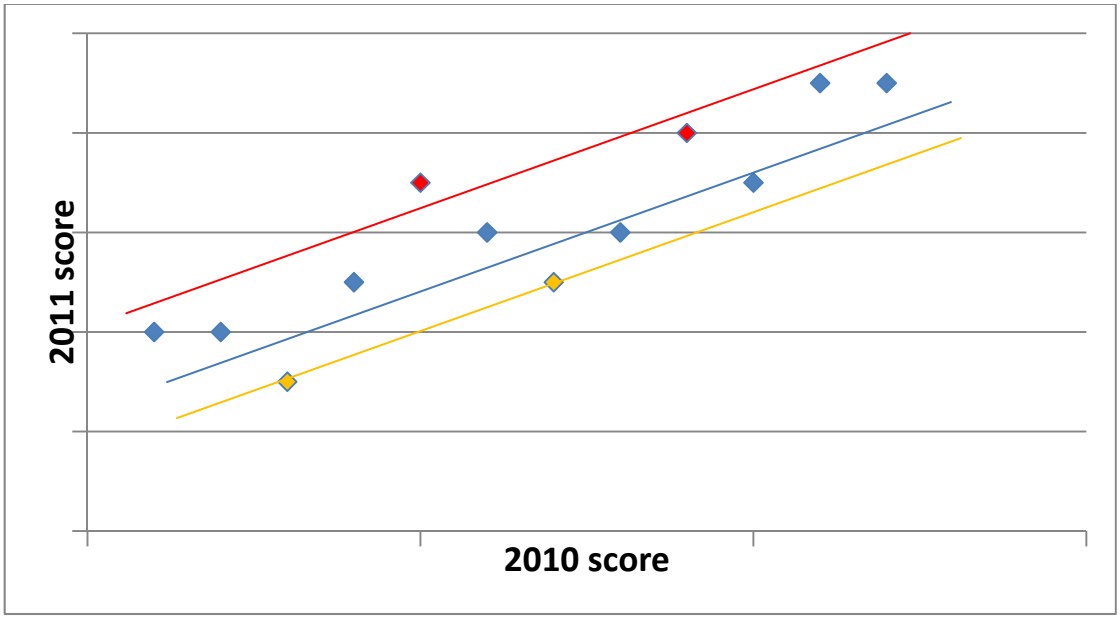


Figure I.3 further refines the predictions, incorporating the additional information about the two students with IEPs. Both of the students with IEPs score below the original blue prediction line, suggesting that students with IEPs do slightly worse in 2011 than non-IEP students who had the same 2010 scores. If we were predicting the 2011 score of another IEP student, we would adjust our prediction downward from the blue line by an amount that is approximately the average distance between the yellow diamonds and the blue line. This adjusted predicted line for students with IEPs is represented in yellow.

Figure I.3. Prediction Based on Last Year's Score + Gifted + IEP



The VAMs implicitly make predictions for every student in a class (or school), using data from across the district (or state) and a wide range of predictor variables (which we describe in detail in Chapter II). Combined, these predictions tell us how any particular class would do if served by the average teacher. Teachers whose classes exceed their predicted scores are above average in value-added terms. Teachers whose classes fall short of their predicted scores are below average in value-added. Note that a value-added result does not provide information about whether a teacher's contribution to student achievement is "good enough." It is inherently a relative rather than absolute measure of a teacher's contribution

Each VAM estimate is reported to teachers or schools as a percentile ranking. For teachers, the percentile ranking estimates where they stand in the distribution of teachers teaching the same subjects and grades *within PPS*. For schools, in contrast, in most cases we report a percentile ranking that estimates where they stand in the distribution of schools serving the same grades *across Pennsylvania*. Ideally, we would use a statewide comparison for all estimates, but many of the student assessments used in the teacher VAMs are conducted only in Pittsburgh, precluding a statewide comparison.

The next three chapters of our report describe the student outcomes that are used in Pittsburgh's VAMs (Chapter II); enumerate the information on students that is used to predict their performance and account for factors outside the control of the teacher or school (Chapter II); discuss the technical details of the VAMs (Chapter III); and explain some of the limitations of the VAMs (Chapter IV). The VAMs used in Pittsburgh are applied not only to test scores on the

Pennsylvania System of School Assessment (PSSA), but also to scores on locally developed Curriculum-Based Assessments (CBA) in a wide range of courses at middle- and high-school levels, to scores on the Preliminary SAT (PSAT), and to student attendance. The VAMs account for factors outside the control of schools and teachers by incorporating statistical adjustments for various student characteristics, most prominently including each student's achievement and attendance in prior years.

The last four chapters of the report explain how VAMs for each student outcome are combined to create a series of composite measures for each school (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); and discusses the application of VAMs for use in two programs designed to recognize and reward outstanding performance: the Promise-Readiness Corps and Students and Teachers Achieving Results (STAR) (Chapter VIII).

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

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

A. Test Outcomes and Baselines

The assessment outcomes that are used in VAM calculations for 2010-11 are catalogued by grade in Table II.1. The outcomes come from the Pennsylvania System of School Assessment (PSSA), from Pittsburgh's curriculum-based assessments (CBA), and from the PSAT.

Table II.1 Available Test Scores by Subject and Grade

Column1	3	4	5	6	7	8	9	10	11	12
PSSA Reading	B	B	B	B	B	B			B	
PSSA Writing			B			B			B	
PSSA Math	B	B	B	B	B	B			S	
PSSA Science		B				B				
CBA Math				B	B	B				
CBA Algebra I							B			
CBA Algebra AB-BC							B			
CBA Geometry								B		
CBA Geometry AB-BC								B		
CBA Algebra II									B	
CBA English				B	B	B	B	B	B	B
CBA African American Literature										B
CBA Earth Science				B						
CBA Life Science					B					
CBA Biology							B			
CBA Chemistry								B		
CBA Physics						B			B	
CBA Civics							B			
CBA World History								B		
CBA US History						B			B	
PSAT Reading								S	S	
PSAT Writing								S	S	
PSAT Math								S	S	

Note: Cells marked with an "B" are test scores used in both the school and teacher VAMs. Cells marked with an "S" are used only in the school VAMs. Tests which are sometimes taken out of grade by students are recorded in the grade cell where the largest number of students take the test.

Most PSSAs and CBAs are used in both teacher and school VAMs, although the PSAT and grade 11 PSSA math are used only for schools, because PPS and the Pittsburgh Federation of Teachers (PFT) have determined that those assessments are not relevant to specific teachers because they are not directly aligned with specific courses. PPS and PFT have chosen not to include the 11th-grade science PSSA in the school VAMs or the teacher VAMs—except, beginning in 2012, for Students and Teachers Achieving Results (STAR) awards, discussed in Chapter VIII. In future years, we will be able to add VAMs for additional assessments such as Keystone exams (now under development by the state) and the grade 3 PSSA, should PPS and PFT decide they merit inclusion.

The validity of using these assessments in VAMs, depends, first of all, on their validity as measures of student learning. Although no standardized assessment can provide a complete and comprehensive picture of everything we expect a student to learn, we assume that the state's accountability tests (PSSAs) are appropriate measures of student learning in the relevant grades and subjects. The district's home-grown CBAs 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 be aligned with any particular courses, but it has been developed and refined by psychometric experts, it has very high year-to-year reliability, and it has been shown to be predictive of preparation for college—one of PPS' key aims.

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 (e.g., Schochet and Chiang, 2010; McCaffrey et al., 2009) has shown that estimates of a teacher's value-added can have a substantial amount of statistical "noise" (i.e., random error) if only one year of teaching is examined. We enhance the reliability of Pittsburgh's VAMs by averaging across multiple years of performance. Whenever possible, VAM estimates for schools are averaged across the last two years and VAM estimates for teachers are averaged across the last three years. Detailed results of the VAM analyses are presented in Chapter VII; they show that the VAMs for every one of these assessments produce results that provide real information in their ranking of schools, rather than simply showing a random distribution.

Table II.2 shows the same list of outcome measures along with the prior test-score measures that are used as baseline controls in each VAM to account for students' own prior achievement. The grade level indicates the grade of the majority of students taking an assessment, but all students taking a particular assessment, regardless of their grade level, are eligible to be included in the VAM analysis. 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 provide additional information about students' baseline knowledge and abilities. All models include at least two tests from prior years. We also accounted for a third prior test in all cases in which adding the third prior test can be done without excluding substantial numbers of students (e.g., those who lack an additional prior test because they transferred into PPS after the particular baseline test was taken).

Table II.2. Assessment Outcomes and Baseline Test Scores Used in PPS VAMs, 2010-11

Outcome	Prior Test 1	Prior Test 2	Prior Test 3
PSSA Math Grade 4	PSSA Math Grade 3	PSSA Reading Grade 3	
PSSA Reading Grade 4	PSSA Reading Grade 3	PSSA Math Grade 3	
PSSA Science Grade 4	PSSA Math Grade 3	PSSA Reading Grade 3	
PSSA Math Grade 5	PSSA Math Grade 4	PSSA Reading Grade 4	PSSA Science Grade 4
PSSA Reading Grade 5	PSSA Reading Grade 4	PSSA Math Grade 4	PSSA Science Grade 4
PSSA Writing Grade 5	PSSA Reading Grade 4	PSSA Math Grade 4	PSSA Science Grade 4
PSSA Math Grade 6	PSSA Math Grade 5	PSSA Reading Grade 5	PSSA Writing Grade 5
PSSA Reading Grade 6	PSSA Reading Grade 5	PSSA Writing Grade 5	PSSA Math Grade 5
CBA Math Grade 6	PSSA Math Grade 5	PSSA Reading Grade 5	PSSA Writing Grade 5
CBA English 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
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 English Grade 7	PSSA Reading Grade 6	PSSA Math Grade 6	
CBA Life Science Grade 7	PSSA Math Grade 6	PSSA Reading Grade 6	
PSSA Math Grade 8	PSSA Math Grade 7	PSSA Reading Grade 7	
PSSA Reading Grade 8	PSSA Reading Grade 7	PSSA Math Grade 7	
PSSA Science Grade 8	PSSA Math Grade 7	PSSA Reading Grade 7	
PSSA Writing Grade 8	PSSA Reading Grade 7	PSSA Math Grade 7	
CBA Math Grade 8	PSSA Math Grade 7	PSSA Reading Grade 7	
CBA English Grade 8	PSSA Reading Grade 7	PSSA Math Grade 7	
CBA Physics Grade 8	PSSA Math Grade 7	PSSA Reading Grade 7	
CBA US History Grade 8	PSSA Reading Grade 7	PSSA Math Grade 7	
CBA Algebra I/AB-BC Grade 9	PSSA Math Grade 8	PSSA Reading Grade 8	PSSA Writing Grade 8
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	PSSA Reading Grade 8	PSSA Writing Grade 8	PSSA Math Grade 8
CBA Geometry/AB-BC Grade	CBA Algebra I/AB-BC Gr 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
PSAT Math Fall Grade 10*	PSSA Math Grade 8	PSSA Reading Grade 8	PSSA Writing Grade 8
PSAT Reading Fall Grade 10*	PSSA Reading Grade 8	PSSA Writing Grade 8	PSSA Math Grade 8
PSAT Writing Fall Grade 10*	PSSA Writing Grade 8	PSSA Reading Grade 8	PSSA Math Grade 8
CBA 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	PSSA Reading Grade 8	PSSA Writing Grade 8
PSAT Math Fall Grade 11*	PSAT Math Fall Grade 10	PSAT Reading Fall Grade 10	PSAT Writing Fall Gr 10
PSAT Reading Fall Grade 11*	PSAT Reading Fall Gr 10	PSAT Writing Fall Grade 10	PSAT Math Fall Grade 10
PSAT Writing Fall Grade 11*	PSAT Writing Fall Gr 10	PSAT Reading Fall Grade 10	PSAT Math Fall Grade 10
PSSA Math Grade 11*	CBA Geometry Grade 10	PSSA Math Grade 8	PSSA Reading Grade 8
PSSA Reading Grade 11	CBA ELA II Grade 10	PSSA Reading Grade 8	PSSA Writing Grade 8
PSSA Writing Grade 11	CBA ELA II Grade 10	PSSA Writing Grade 8	PSSA Reading Grade 8
CBA ELA IV/AA Lit Grade 12	CBA ELA III Grade 11	PSSA Reading Grade 8	PSSA Writing Grade 8

* indicates outcome is included only for schools, not for teachers

B. Non-Test Outcomes and Baselines

Although VAMs typically involve test outcomes, three of the school-level VAMs to be used in Pittsburgh also include non-assessment measures of student outcomes: the passage rate of core courses (in high schools), student attendance (at all grade levels), and holding power, i.e., a school's ability to keep a student in school the next year (in high schools). Including these non-test outcomes offers a more comprehensive view of a school's effect on student achievement which might not be represented by test scores alone. These are not used as outcomes for teacher VAMs because we assume they cannot be attributed to individual teachers. The method of estimation for the non-test outcome VAMs differs slightly from the method used for tests; see Chapter III, part E for details. Table II.3 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 (i.e., current attendance rate and core pass rate are related to previous achievement as well as previous attendance and core pass rates). As with the test measures, the VAMs for attendance rate, core pass rate, and holding power 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, core pass rate, and continued enrollment, after accounting for student characteristics and previous performance. Attendance and core pass rate VAMs involve comparisons only within PPS, because equivalent data are not available statewide; the holding power VAM uses a statewide reference. Holding power is not yet reported in school VAM reports, but is currently intended to be used beginning in 2012 as part of the STAR awards system (described in Chapter VIII).

Table II.3. Non-Assessment Outcomes and Baseline Measures Used in School VAMs, 2010-11

Outcome	Grade(s)	Baseline Measures	Baseline Test, Grade(s)
Attendance rate	1-3	Prior attendance rate, K-2	None
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
Holding power {0,1}	9-11	None	PSSA math & PSSA reading & PSSA writing, 8

Note: At the elementary level, baseline PSSA math and reading scores are available only in grade 4 and 5. Attendance VAMs in earlier grades do not include baseline measures for prior achievement. Prior core pass rate also include indicators for perfect attendance and 100 percent prior pass rate.

1. Core Course Pass Rate

The VAM using pass rates for core courses is designed to provide useful information on how effective a school is at moving students toward a high school diploma, accounting for the prior progress of the same students. Core courses are defined to be those in math, reading/language arts, science, and social studies. This currently includes all courses related to those subjects, but in future years (and in the forthcoming calculations of 2011 Promise-Readiness Corps results), the definition of "core" may be narrowed so that it excludes courses that are related to these subjects but are not the primary courses in the subjects (e.g., Yearbook as a course in English). Using Pittsburgh's course data, we determine the percentage of core courses that a student passes, and then we apply a value-added model to that percentage (i.e., 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 the percentage. However, in the Pittsburgh data, the variable that tracks the number of courses/credits a student needs to graduate is not measured consistently across all high schools. Because of curriculum differences across high schools, some students have access to a different number of core courses than other students in the district. Using the percentage of core courses passed allows us to account for these curriculum differences. If PPS begins to use the core pass rate for evaluation purposes, it will be important for the district to monitor and ensure that standards for passing a class are maintained, because making pass rate an accountability measure will create incentives to lower the standards for passing in order to bolster a school's score.

The value-added estimate for core-course pass rate is not included in the school VAM reports for 2010-11. It is, however, included in the VAM estimates for Promise Readiness Corps teams at the high schools (discussed in Chapter VIII).

2. Attendance Rate

For grades 1 through 12, we estimate schools' contributions to students' rates of attendance during the school year, accounting for their attendance in the prior year. In constructing our measure of attendance, we do not distinguish between excused and unexcused absences. Although excused and unexcused absences are given separate codes in Pittsburgh's data, our examination of those data suggested that standards for determining whether an absence is excused or unexcused may vary over time and among schools. Overall absence rates are more stable over time and across schools than are rates of excused or unexcused absences. We therefore use the overall attendance rates in the current VAMs. As with the 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 own attendance rates in the preceding year.

In future years, we recommend that grades 1-3 be dropped from the attendance VAMs. After estimating school value-added for attendance, we discovered that the estimates for grades 1-3 are far less precise than those for higher grades. The VAM for grades 1-3 includes a control for prior attendance but not for prior test scores, which do not exist for students that young. In contrast, prior attendance and prior test score controls are available and used in the VAMs for grades 4-8 and 9-12.[1] At higher grade levels there is more variation in school effectiveness in terms of student attendance and more precision in the estimates. The lack of prior achievement controls in grades 1-3 is likely to be an important factor in explaining low overall precision, and we expect that precision would improve for K-5 and K-8 schools if grades 1-3 were excluded from future attendance rate VAMs.

The attendance rate VAM, like the core-course pass rate VAM, could produce incentive problems. Specifically, if value added for attendance is used as an outcome for evaluation purposes, incentives arise for schools to mark fewer students absent each year. We strongly recommend that

[1] We controlled for prior-grade scores in the VAM for grades 4-8 and eighth grade scores in the VAM for grades 9-12.

PPS monitor and ensure that standards for recording attendance remain consistent over time and across schools.

3. Holding Power

For students in grades 9 through 11, we will measure the extent to which Pittsburgh high schools are successful at keeping students enrolled in school. This measure does not examine whether students remain enrolled in the same school or even in the district, but in any school in the state (assuming they have not graduated). We define a student as "held" if the student enrolled in a Pennsylvania public school the following year (or graduated). To obtain this measure, we will use statewide data from the Pennsylvania Department of Education. Using the statewide data allows the holding power measure to account for inter-district mobility within Pennsylvania.

Whether a student is "held" is the raw measure, but like all of the other measures described here, the application of a VAM to the raw measure aims to estimate the school's contribution rather than merely comparing raw numbers. The holding-power VAM estimate accounts for 8th-grade PSSA scores and other student characteristics to make its predictions. It is measured relative to the statewide average in terms of a school's ability to keep students enrolled at rates higher than predicted for those students.

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.4 defines the student background characteristics that are included in all teacher and school VAMs. Some VAMs include additional background variables as well. We describe these cases in the context of their specific VAM applications.

In addition to student characteristics, we also account for peer influences in most models. The peer measures we use are for the following characteristics: gender, meals program, English language learner status, gifted status, disability rate, prior year absence rate, prior year suspension rate, prior year full district membership, and prior year average PSSA math and reading scores.[2] The variables indicate the average rate of a characteristic across youth enrolled in the student's classroom. When a student takes multiple courses during the year in a subject, the peer variables are averaged across classrooms.

Teacher VAMs also account for class size, which is presumably not under the control of teachers. Class size is not included as a variable in the school VAMs, however, because schools may have some influence over class size.

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

Table II.4. Variables for Student Background Characteristics in Pittsburgh Teacher and School VAMs, 2010-11

Background Variable	Definition
Male	Male gender
Meals program	Free or reduced price meal eligibility status
Race/ethnicity	African-American, white, Asian, Hispanic
English language learner	English language learner status
Gifted	Participation in the gifted program
Pittsburgh Scholars Program	Taking a class in the Pittsburgh Scholars Program
Advanced Placement	Taking an advanced placement class
Center for Advanced Studies	Taking a Center for Advanced Studies class
Specific learning disability	SLD designation under Individuals with Disabilities Education Act (IDEA)
Speech or language impairment	SLI designation under IDEA
Emotional disturbance	ED designation under IDEA
Mental retardation	MR designation under IDEA
Autism	AUT designation under IDEA
Physical / sensory impairment	An IDEA designation for hearing impairment, visual impairment, deaf-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 2010-2011 school year
Grade repeater	Repetition of the current grade
PSSA-Modified	Student took modified version of the PSSA (PSSA outcomes only)
Absence rate (prior year)	Prior year absences divided by days of enrollment
Suspension rate (prior year)	Prior year days suspended out-of-school or expelled divided by days of enrollment
Full year district membership (prior year)	Enrolled the entire prior school year in Pittsburgh
Magnet applicant	Has applied for entry to a magnet program
Age	Student age in years as of the beginning of an academic year (September 1) including fractional years of age
Behind grade for age	Student age is 1.5 years older than typical for grade level
Special Services	Applied for special services, such as attending a school outside of feeder pattern or special education services

Notes: All variables are binary with the exceptions of absence rate, suspension rate, and age. We aggregate several of the lowest incidence disabilities because some individual categories do not comprise even a single student at all grade levels.

Pittsburgh has three main types of advanced courses at the high school level: Pittsburgh Scholars Program, Center for Advanced Studies, and Advanced Placement.[3] Because course selections are made at the beginning of the school year, they are outside the control of current-year teachers. We therefore add indicator variables for each advanced course type to the teacher VAMs to protect the effect estimates from being biased upward for teachers with more advanced students. We omit the course-type variables from the school VAMs, however, because schools may be able to influence the availability of these programs or the amount of resources designated to them.

We also received data from PPS indicating whether a parent successfully requested special services (such as attending a school outside their feeder pattern or special education services). In addition, we received a variable that indicates whether students ever entered a magnet school lottery (whether or not the application was successful). We include these variables in the value added models to help control for unobserved motivation and effort levels of students and their parents that can influence achievement growth.

Many of these background variables are in fact related to student achievement in Pittsburgh (consistent with many published studies). Not surprisingly, students' own prior achievement scores have the largest relationship to their current academic performance. Student-level characteristics like age, gender, race, meals program, disability, and gifted indicators tend to be statistically significant as well, even controlling for other observable factors. Among these variables, gifted status tends to have the largest magnitude. We find some evidence that absences relate negatively to score gains. Course type, magnet application, and class size do not show consistent patterns, after controlling for other factors. The class average variables tend not to be statistically significant.

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

III. TECHNICAL DETAILS OF PPS VALUE-ADDED MODELS

A. Detailed Value-Added Model Description

The following general statistical equation describes the VAMs:

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

In the model, $Y_{i,t}$ is the outcome for student *i* in year *t*. The models are estimated separately for each grade, subject, and assessment. For example, $Y_{i,t}$ could be a student's score on the grade 5 math PSSA during 2010-11. $B_{i,t-1}$ is a vector of baseline scores for student *i* from a prior year to account for students' own academic histories. The baseline scores typically come from the previous school year, although baseline scores can come from up to three prior years at the high school level depending on the availability of assessments in consecutive grades. We include two or three baseline scores rather than one because all assessments measure students' content knowledge with some degree of error, which can bias estimates if not adequately addressed. This also helps mitigate potential issues related to test ceiling effects. Whenever possible, at least one baseline score comes from the same subject area as the outcome measure.[4] See Table II.2 for the full list of outcome measures and baseline variables.

Students with baseline scores near the top or bottom of the distribution might experience different trajectories than students who scored near the middle of the distribution. We account for this possibility by including a quadratic polynomial function. This helps to account for the possibility that students who are near the top of the distribution may not be expected to show as much growth relative to students who are closer to the middle of the distribution.

$X_{i,t}$ is a set of variables for observable student characteristics, while $\bar{X}_{i,t}$ represents observable peer characteristics (described for Pittsburgh VAMs in Table II.3 and the surrounding text[5]). $D_{i,t}$ is a set of variables for a student's teachers in a subject or schools during the year, T is a set of school year indicators, and $e_{i,t}$ 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 VAM. Each δ coefficient identifies a teacher's contribution or a school's contribution to student learning—the extent to which the actual achievement of students tends to be above or below what is expected for the average teacher or school.

We define the average VAM score (i.e., the average δ coefficient) to be a zero value, but this does not mean that student learning is zero for the teacher or school with the average VAM score. Rather, it means that positive VAM estimates represent above-average (above predicted)

[4] An example of an exception would be the grade 4 science PSSA, where a same-subject baseline score is not available. Both baseline scores would come from other subjects (e.g. grade 3 math and reading PSSAs).

[5] Some of the $X_{i,t,y}$ variables are correlated with each other. Including related variables in VAMs does not mean that teacher or school effects will be estimated inconsistently. In fact, it typically improves the validity of VAM estimates so long as both of the related variables are relevant to student achievement growth. We correlated each of these control variables with each other and did not find any correlations that seemed unreasonably large or surprising in their direction.

teacher/school performance and negative VAM estimates represent below-average (below predicted) teacher/school performance. For schoolwide, test-based VAMs and holding power, average performance is defined at the state level by creating a hypothetical distribution of statewide performance as described in Chapter VI. For attendance and core-course passing VAMs and all teacher VAMs, average performance is defined among Pittsburgh schools and Pittsburgh teachers.

The VAMs for teachers and for schools differ in the number of cohorts of student data they include. VAMs for schools examine two years of teaching, producing an average of a school's performance across the 2009-10 and 2010-11 school years. VAMs for teachers examine up to three years of teaching, producing an average of the teacher's performance across the last three school years. In some instances, insufficient historical data is available to include three years of teaching. Multi-year VAM estimates are less prone to random and systematic fluctuations that stem from being assigned a few students who end up displaying unusually high or low achievement growth. (Information on an exploratory analysis that compares the precision of single-year VAM estimates in Pittsburgh with that of three-year VAM estimates can be found in Appendix Table A-1.) They can therefore detect performance differences with greater validity and reliability, which is advantageous for any high-stakes application. However, multi-cohort VAMs are less reflective of immediate past performance because they average annual value-added scores over multiple years.

B. Standardization

Because VAM estimates reported in assessment units (e.g. PSSA scaled score points) are not comparable across tests, grades, subjects, or years, we standardize all outcome measures (test-based and non-test-based) prior to running the analyses. Specifically, we map assessment units to a standard measure, called a z-score, by subtracting the average value (e.g., the average grade 4 math PSSA scaled score) from individual scores by school year and then dividing by the standard deviation of scores.[6] Expressing scores like this allows us to interpret above-average scores in terms of how close to average most students tended to fall, regardless of the assessment. Similarly, we standardize the data for all baseline/background variables based on the analysis sample for each VAM and exclude the constant term. This latter standardization process enhances precision, and the exclusion of the constant term means that the teacher and school effects will be measured relative to the average contributions of teachers and schools.

C. Teacher and School Dosage

All of our models for Pittsburgh use a dosage approach that allows the VAM to account for the fact that some students change teachers or schools during the school year. Specifically, a dosage approach accounts for the extent to which students are exposed to different teachers and schools during the school year. It provides a finer level of detail than simply tracking with a binary indicator whether a student was taught at a given school or by a given teacher at all during the year. For example, if a student moves schools in Pittsburgh during the year and is enrolled in one math class while at each school, the teacher and school dosage values in the VAM would be fractions between 0 and 1 based on the days enrolled at each school. We split teacher dosage values evenly when students take multiple courses in the same subject at the same school or (to the extent identified in

[6] A standard deviation measures score variation—what we can see graphically by whether the distribution of scores tends to be spread out or grouped tightly together.

the data) appear to have two teachers in the same classroom. We account for time not enrolled or enrolled outside the district by including a "residual dosage" term that equals one minus the sum of dosage values across teachers or schools.

We estimate value-added for teachers only if they are identified as teachers of the primary course in the relevant subject. This means, for example, that even though Yearbook is identified in PPS data as a course in English, the instructor for Yearbook is not given a value-added score in English. Only the student's primary English teacher for that grade receives an English VAM estimate.

The validity of the use of a dosage method to account for student transfers between classrooms and across schools depends on accurate course and school roster information. In future years PPS is planning on using a roster verification system to ensure the accuracy of the dosage data used in the VAMs.

D. Shrinkage

VAMs typically use a procedure known as empirical Bayes estimation or shrinkage to address the fact that, among teachers/schools with the same level of true performance, those with fewer students in the estimation sample face a greater likelihood that their students happen, by chance, to have atypically high or low learning growth driven by other factors.[7] In the absence of a shrinkage adjustment, teachers with fewer students—that is, those with less precise estimates—will tend to be overrepresented at both the high and low ends of the estimated performance distribution just by chance. Shrinkage adjustments account for the fact that estimates with greater precision carry greater strength of information about teachers' true performance levels. The adjusted estimate is a weighted average of the individual initial estimate and the mean estimate across teachers, with more precise initial estimates receiving greater weight. In essence, teachers are assumed to be average in performance until evidence justifies a different conclusion. To further minimize the risk of making erroneous conclusions on the basis of imprecise estimates, we limit analyses to teachers who taught more than 10 students during the year. This type of restriction, common in the research literature, reduces the potential for teacher effects to be influenced just by the scores of one or two students (Kane and Staiger, 2002; McCaffrey et al., 2009). We use this same process in school models as well, but it is less important than in teacher models because sample sizes are larger.

E. Technical Details for VAMs for Non-Test Outcomes

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

[7] The shrinkage procedure is an empirical Bayes procedure based on Morris (1983) that minimizes the mean squared error of the value-added estimates.

- **Core Pass Rate**: Approximately 65% of PPS high school students in 2009-10 passed all their core classes, which means that they reached the upper limit of the core pass rate metric. In situations like this, ordinary linear regression models like Equation (1) provide biased estimates, because the relationship between higher values for the background variables and a higher core course pass rate becomes nonlinear when students pass all courses. To eliminate this bias, we use a Tobit model to estimate the VAM. 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 equaling one if the student had a perfect pass rate in the prior year.
- **Attendance Rate:** Because approximately 5 percent of students have perfect attendance in any given grade, we estimate this VAM using the Tobit model as well. 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. Because assessment data are not available before grade 3, attendance rate VAMs for grades 1 through 3 do not account for prior academic achievement.

IV. METHODOLOGICAL LIMITATIONS

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

A. Non-Random Assignment of Students

Students are not randomly assigned to teachers and schools, which may introduce bias to any VAM estimates if not adequately addressed (Rothstein, 2010). Chaplin and Goldhaber (2012) find, however, that the sorting bias may be small relative to the variance of teacher value-added scores. Our VAMs assume that assignment is functionally random once we account for the observable factors included in the VAMs. Even though this assumption may not be strictly true, research evidence suggests that resulting bias in the VAM estimates is likely to be small. Kane and Staiger (2008) found that variation in teacher VAM scores significantly predicted achievement differences in a subsequent year when classrooms were assigned randomly. This suggests that any bias that exists does not prevent the VAMs from identifying an important component of school and teacher performance.

B. Distinguishing Between School and Teacher Effects

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

An alternative method to account for the influences of schools in teacher level VAMs would be to add variables accounting for school characteristics. We cannot adjust for most school characteristics that might be directly relevant to teacher value-added (e.g., resources available, principal quality, school safety), because data are not readily available on those characteristics. Even if this data were readily available, it is difficult to separate the effect of a school having good characteristics from the possibility that good teachers choose to work at schools with attractive characteristics; doing so requires variation in characteristics for the same school over time and substantial transfer of teachers across schools. Nonetheless, we performed exploratory analyses that included measures that are available in Pittsburgh's data and that might serve as proxies for school-level characteristics that could affect teacher value-added. These proxy variables included schoolwide averages of the number of days students are suspended, percent of students eligible for free or reduced price meals, and prior student test scores. The exploratory analyses found that these school characteristics explained very little of the variation in student outcomes. More to the point, the teacher value-added estimates were almost identical with or without the inclusion of the school characteristics. The lack of explanatory power of these variables could be due to their being poor proxies for actual factors affecting teacher effectiveness or to small variation in these characteristics 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 VAMs are nearly identical in their analytic structure—differing only in whether teacher or school dosage variables are used—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 (e.g., 6 and 9), students are generally served by the same school both in the current year (i.e., the year to which a set of VAM estimates apply) and in the prior year, when baseline scores are measured. This implies that some variables that we assume are outside the control of a school for a current year value-added model were actually affected by that school in the prior year. For example, a school could hold a student back for a grade, which would affect next year's VAMs differently than if the school had allowed the student to progress to the next grade. This is not typically an issue in teacher models, because students generally change teachers each year (except in rare instances when teachers "loop" to the next grade with their students, in which case the VAM operates like a school-level VAM).

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

D. Absence of VAM Estimates for Grades K-3

Because reliable baseline test measures do not yet exist for grades K-2, we cannot include the first four years of schooling (grades K-3) in the VAMs for 2010-11. In 2010-11, PPS introduced the Terra Nova test in grade 2, which will make it possible next year to estimate a VAM using the grade 3 PSSA as the outcome and the Terra Nova as the baseline. Starting in 2012-13, PPS plans to introduce the Terra Nova to grades K and 1 as well. Although expanding VAM coverage to lower grades will be feasible in future years, for now we must assume that school VAM estimates covering grades 4 and 5 are reasonable proxies for a school's performance in grades K-5.

We conducted exploratory work in assessing value-added from grades K to 3 using the Dynamic Indicators of Basic Early Literacy Skills (DIBELS). DIBELS is a set of diagnostic tests designed to be administered to students between kindergarten and 6th grade to assess their progress in learning to read. We have data on DIBELS scores from 2004 to 2009 in Pittsburgh for students between kindergarten and 3rd grade. We used this data to run exploratory school-level VAMs that assess the effectiveness of schools at improving students' reading abilities through the 3rd grade.

We used 3rd grade PSSA scores as outcome variables and beginning of kindergarten DIBELS scores as the baseline scores. We used beginning of kindergarten tests as the baseline measures to obtain a measure of student ability before the school had a chance to influence the student, so that

gains represented the cumulative effect of the school during the initial years of schooling. The beginning of kindergarten DIBELS diagnostics had almost no explanatory power in these VAMs, however, meaning that the school effect estimates were almost the same whether or not we included the prior DIBELS scores. This lack of explanatory power causes us to be concerned that kindergarten DIBELS diagnostics do not provide adequate indicators of student ability. We then explored the possibility of using DIBELS results from grades 1 and 2 as additional baseline variables. While these measures provided some additional explanatory power, they were still not as highly correlated as tests usually are with prior-year scores in other grades. In addition, once we use DIBELS scores from grades 1 and 2 as baselines, we no longer could estimate a school's cumulative effect from kindergarten through 3rd grade.

An additional potential problem with using DIBELS data is that the diagnostics are administered to students individually and scored by teachers. Using DIBELS for teacher and school evaluation purposes could create a conflict of interest for teachers scoring the tests. Due to the lack of explanatory power of the kindergarten DIBELS diagnostics and the potential incentive problems surrounding future DIBELS tests, we do not recommend using it for value-added purposes.

E. Missing and Omitted Data

VAMs can account only for factors that are measured in the data, meaning that estimates may be biased if important background/baseline variables cannot be included.

1. Missing Data on Resources for Students and Schools

Pittsburgh's data system does not currently track student participation in some intervention programs (during the school year or in summer school) that may help raise test scores. Any effect of these programs will be mistakenly attributed to the classroom teacher or school of record. Because PPS plans to begin collecting centralized data on participation in interventions, VAM estimates in future years should be able to factor out the effects of the interventions.

We also lack information on the number of hours of instruction in particular subjects, meaning that we cannot account for some teachers or schools spending more time, for example, on social studies than do other teachers or schools. Ignoring differences in instructional hours or available resources could be problematic to the extent that they are outside the control of a teacher or school. We do not, for example, adjust results for Pittsburgh's Accelerated Learning Academies, which have more total instructional time than other schools.

2. Missing Baseline Test Scores

In each VAM, between 5 percent and 10 percent of the students that have data on the outcome measure are dropped because of missing data on at least one baseline/background variable. In most cases, the missing element is a prior test score. Prior scores can be missing for several reasons, such as when students transfer into Pittsburgh from outside the district, take a test out of grade, or are absent from school during testing in the prior year. To increase precision, it is possible to impute the missing prior test scores for these students and thus keep them in the VAM. 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 VAM. However, it can be difficult to find a previous test score that can be used to impute the missing values consistently for each VAM. Also, imputation is not feasible for students who transfer to Pittsburgh from other districts because there is no information in

Pittsburgh's data collection on any of their prior scores. Therefore (with the exception described below) we do not impute missing values.

Missing data tends not to be a persistent problem for most students over time. For example, if a student transfers from another district in 2009-10 and is missing prior test score data he or she will take the normal end-of-year assessments in Pittsburgh. That student would be dropped from the 2009-10 VAMs due to missing prior test score data, but will appear in the 2010-11 VAMs, because the end-of-year assessments in 2009-10 could be used as the baseline scores needed in the 2010-11 VAMs. Therefore, except in cases of rapid mobility, students tend to be picked up by the VAMs after a full year in the district.

3. Substituting 9th-Grade Entry SRI Scores for Missing 8th-Grade PSSA Scores

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

To prevent the exclusion of a relatively large fraction of students from the high school VAMs, we impute values for missing 8th grade PSSA scores using data on student achievement at the beginning of 9th grade. On average, the imputation increases our VAM sample sizes by a little more than 10%. Starting in 2010-11, the Scholastic Reading Inventory (SRI) was given to all PPS students in September of 9th grade. We use the initial 9th grade SRI in 2010-11 to impute the values for missing 8th grade PSSA reading, writing, math, and science scores for students in 2009-10.[9] The 9th-grade SRI score gives us a way to estimate what the student's 8th-grade PSSA score would have been.[10] Performing this imputation may be important to avoid bias in the results, because the number of 9th-grade students who are missing 8th-grade PSSA scores varies widely across Pittsburgh's high schools.

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

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

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

V. COMPOSITE VALUE-ADDED MEASURES

Every school in Pittsburgh has VAM results for at least five different outcome measures; high schools have many more. Many teachers likewise have VAM estimates related to more than one student assessment. At the policy direction of PPS and the PFT, the VAM results for the individual test-based outcomes are aggregated into composite measures for reporting purposes and for informing awards for Promise-Readiness Corps teams and STAR schools (both described in Chapter VIII). This creates a simpler presentation of results and allows educators to get a sense of a school's subject-wide and overall performance. Composite measures are obtained by subject (for schools only) and across all test-based measures (for schools and teachers). For example, the math composite for a middle school incorporates VAM data from PSSAs and CBAs in grades 6, 7, and 8.

A. Composition of Composites

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

Table V.1. The Composition of Subject Composites for Pittsburgh School VAMs, 2010-11

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

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

Fourth-grade PSSA science and 8th-grade CBA US History are not reported for K-5 and 6-8 schools because some schools have only one teacher for these subjects; reporting results for individual assessments would therefore implicitly identify a teacher. They are, however, included in

subject composites for K-8 and 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 each school in all relevant grade configurations.

Because many teachers receive VAM estimates based only on one or two assessments in the same subject, subject level composites are not reported for teachers. Instead the VAM score for each assessment is reported, along with an overall composite estimate that includes all of a teacher's relevant scores. The overall composite estimate for a teacher implicitly compares the teacher to all other PPS teachers whose students take at least one of the same assessments.

B. Construction of Composite Estimates: Precision Weighting

The composite measures are obtained by combining the individual VAM estimates using a two-step method.[11] In the first step we normalize the individual VAM distributions so they each have the same standard deviation. The standard deviation of a distribution is a measure of its spread—how concentrated or spread out it is relative to the average value. Prior value-added studies, including Mathematica analyses using Pittsburgh 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 may vary across grades within a subject too. When not normalized, a simple average of VAM scores (e.g., an average of a school's VAM scores in grade 4 and 5 based on the math PSSA) will implicitly give more weight to the distribution with the larger standard deviation. For example, the VAM score of a top-performing school according to the measure with the larger standard deviation (e.g., a school scoring at the 95th percentile) will be further away from the average value, and thus larger, than the VAM score of a top-performing school according to the other measure. By normalizing the VAM distributions, we allow for a simple averaging of scores to weight each measure equally.

Rather than combining measures into a simple average, in the second step, we combine measures based on the precision with which they are estimated. All estimates (including observation-based measures of teacher performance as well as value-added measures) are measured with some amount of uncertainty. The uncertainty stems both from the finite number of students included in each VAM and from the statistical noise (i.e., random error) with which each assessment measures student achievement. Statistical noise is the random variation in student test scores that can originate from numerous potential sources, such a student who happens to know a lot about the writing prompt on an assessment, or a cold that is having a lingering effect on several students in the class on test day. Precision is enhanced by more students taking an assessment, because there is more information available to use in measuring performance. Precision is reduced by statistical noise, because it is more difficult to discern whether an increase in a student's test score is due to the performance of the teacher or school versus something else. VAM estimates tend to be noisier in subject areas where it is more difficult to measure student achievement.

The standard error of a value-added estimate indicates the size of the confidence interval—the band of neighboring values around an estimate that is statistically indistinguishable from the estimate. Smaller standard errors indicate greater precision for value-added measures because the

[11] Standard errors for the composite measures are constructed using a modified version of the approximation suggested in Isenberg and Hock (2010).

interval of neighboring and statistically indistinguishable values is also smaller. Precision weighting uses the data to determine which VAMs provide the least noisy estimates of teacher value-added or school value-added and puts more weight on these outcomes.[12] By making use of this information, the method produces a composite that maximizes precision.

Precision weighting involves two trade-offs. First, it gives more weight to some grades and subjects than to others. If reading scores tend to be noisier than math scores, for example, they will contribute less weight to the composite. Second, precision weighting does not capture the views of educators and policymakers about the relative importance of different outcome measures. The student assessment that produces the most precise VAM estimates may not be the one that is most important for long-term success, or the one that receives the largest amount of instructional time. The relative importance of different student assessments could justifiably lead PPS and the PFT to choose weights that differ from the precision-maximizing weights in the future. This would produce a composite measure that has more statistical noise than a precision-weighted composite, but that might better reflect Pittsburgh's educational goals. In fact, as we discuss in Chapter VIII, for some applications the VAM composites are informed by precision weights but adjusted to account for the policy preferences of PPS.

[12] For each assessment we use the inverse of the average variance over all the teacher (school) value-added estimates using to determine the weight that assessment receives in the teacher (school) composite estimate.

THIS PAGE LEFT BLANK FOR DOUBLE-SIDED PRINTING

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

PPS seeks value-added measures that allow comparisons of schools to other schools in the state of Pennsylvania. Results produced by VAMs are inherently relative to the teachers/schools included in the full dataset. Conducting VAMs with statewide data (rather than only within-Pittsburgh data) therefore has the advantage that it allows us to observe 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, is not as rich as Pittsburgh's own data. Most importantly, Pittsburgh has data on many student outcomes that are not available statewide, including CBA results, attendance, and progress in completing core courses. In addition, Pittsburgh has more data on students that can be used to improve the predictions of their likely performance. Relying exclusively on statewide data would therefore dramatically reduce the number of assessments that could be used in the VAMs, and would reduce the overall quality of the analyses.

Instead, we have adopted a hybrid approach for the school-level VAM estimates that capitalizes on the breadth of the statewide data and the richness of Pittsburgh's local data, running VAMs separately but in parallel on both data sets. To make the most of the data from district and state sources, we conduct within-district VAMs to produce fine-grained assessments of how PPS schools perform relative to each other, and we use statewide VAMs to assess where the district as a whole falls in the statewide distribution of performance. This produces a crosswalk or indirect comparison of VAM scores that allows us to estimate the performance of each PPS school relative to the statewide average, without discarding the richer information included in the district's own data.

For teachers, in contrast, we 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.[13] While all schools in Pittsburgh have data on at least one assessment that is given statewide, many teachers can be assessed only with CBAs, none of which are available statewide.

Placing Pittsburgh schools in the statewide distribution involves two steps, which will be discussed in turn. In the first step, we use data on students across Pennsylvania to estimate statewide school VAMs based on PSSA exams. This is done for individual assessments and for composites. Second, we assign a statewide percentile to each Pittsburgh school based on (a) the distribution of Pittsburgh schools in a corresponding composite statewide VAM, and (b) the finer-grained VAM rankings produced using the district-specific data. In other words, if the statewide analysis of PSSA data tells us that the top-performing Pittsburgh middle school in math is at the 95th percentile of statewide value-added, then the Pittsburgh school that we identify as top-performing *in the district based on PSSAs and CBAs* is assigned to the 95th percentile of the statewide distribution. Depending on how well it did on math CBAs, that school might or might not be the same school that landed at the 95th percentile based only on math PSSA scores.

This process accomplishes the dual goals of PPS that school value-added be reported in the context of state performance while also incorporating assessments like CBAs that are offered only in

[13] For example, since the only statewide assessment available at the high school level is the 11th grade PSSA, mapping teacher performance to a statewide distribution would require the assumption that the placement of 9th and 10th grade Pittsburgh teachers in Pennsylvania is the same as the distribution of 11th grade Pittsburgh teachers in Pennsylvania.

Pittsburgh. Through the latter goal, we incorporate information on additional measures that are closely tied to the actual curriculum and cover a broader set of grades and subjects than could be covered by statewide assessments alone. The two-step process also allows us to make use of finer-grained background variables available in Pittsburgh that are not available statewide.

VAM measures for Pittsburgh schools include both state and locally-developed 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 Pittsburgh schools in the statewide VAM distribution as measured by PSSA scores is a reasonable proxy for how they would place if all outcomes in the same subject were available statewide (i.e., if the rest of the state had CBA results alongside PSSA results). The rest of the chapter describes the two-step process in more depth.

A. Statewide Teacher and School VAMs

Using student data from the Pennsylvania Department of Education (PDE), we estimate statewide VAMs that resemble those described in the preceding chapters as closely as possible given the available data.[14] That is, the statewide VAMs involve similar data elements and contain the same features like 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 only include assessments and other dependent measures for which data exist across the state, i.e., PSSAs. Reliable state data do not yet exist on student attendance or core course passage, CBAs are not administered outside of PPS, and the PSAT is given only to a subset of students outside of PPS.
- **Sample includes more students.** The sample size for each statewide VAM is substantially more than it is when estimating a VAM based on Pittsburgh 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 will lead to more precise value-added estimates.
- **Differences in student-level control variables.** The state data contain much of the same student information that we use in the PPS VAMs, although 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, and prior full year district membership.[15] To limit the potential bias associated with including fewer

[14] Assessment data come from the Bureau of Assessment and Accountability (BAA). All other student data come from the Pennsylvania Information Management System (PIMS).

[15] Student attendance is a field in the PIMS but the data are not yet being released by PDE. We exclude the gifted program participation field in PIMS from the state VAM analyses because we are concerned about its validity. The data suggest that no Pittsburgh students participate in the gifted program. In contrast, Pittsburgh's own data indicates that Pittsburgh students participate in the gifted program at a rate that is double the statewide average. We opted against replacing PIMS gifted data for PPS with RTI information because we are concerned with the validity of gifted information in other districts as well.

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).

- **Less exact dosage measure**. Because data on mid-year student transfers are not currently available at the statewide level, school dosage measures are less exact in the statewide VAMs than in the PPS VAMs. We determine the number of schools a student attended during the year and assume an equal dosage between them. We cannot include residual dosage terms because we do not know students' enrollment periods, meaning that we must assume that students are enrolled in a Pennsylvania school the entire school year.

B. Assigning a State Value-Added Percentile to Results Based on Pittsburgh Data

Based on how the distribution of performance in Pittsburgh falls relative to the state, the final step is to assign a state value-added percentile to the Pittsburgh VAM estimates. The distribution of PPS-specific VAM estimates 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 VAM estimates are used to determine the general ranking of PPS schools' performance in the state, and PPS-specific VAM estimates use finer-grained data—including more student-level variables and additional outcomes—to provide a better indication of where each PPS school falls in the district wide distribution.

All schools, regardless of grade configuration, are placed in the same distribution when determining the statewide percentile rank. This means that a school's percentile rank is relative to all schools in Pennsylvania. Since each VAM is estimated separately by assessment and grade and as a result of the normalization of VAM estimates described in Section V.B, a school's place in the statewide distribution 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 Pittsburgh schools (K-5, K-8, 6-8, 6-12, and 9-12) preclude the comparison of schools only to other schools with the same grade configurations when determining the percentile rank. It is therefore necessary to place schools into one statewide distribution to ensure that all schools with overlapping grade ranges are compared to each other.

Figures VI.1 and VI.2 illustrate where PPS schools fall in the statewide distribution on the overall composite value-added measure. More Pittsburgh schools are below the statewide median in terms of overall value-added, although several schools far exceed the median. As indicated in Figure VI.2, the median Pittsburgh school places at the 36th percentile in the statewide distribution using data from 2009 to 2011. The range of composite value-added scores across Pittsburgh schools ranges from the 6th to the 98th percentile, spanning nearly the entire statewide distribution. Red horizontal lines in the figure indicate the 75th and 85th percentiles of the statewide composite value-added measure to highlight Pittsburgh schools that perform especially well relative to other schools in the state. Eight Pittsburgh schools place in the top 25 percent of schools statewide and three schools place in the top 15 percent of schools statewide.

Figure VI.1. Distribution of Composite School VAM Estimates in Pennsylvania, 2009-11

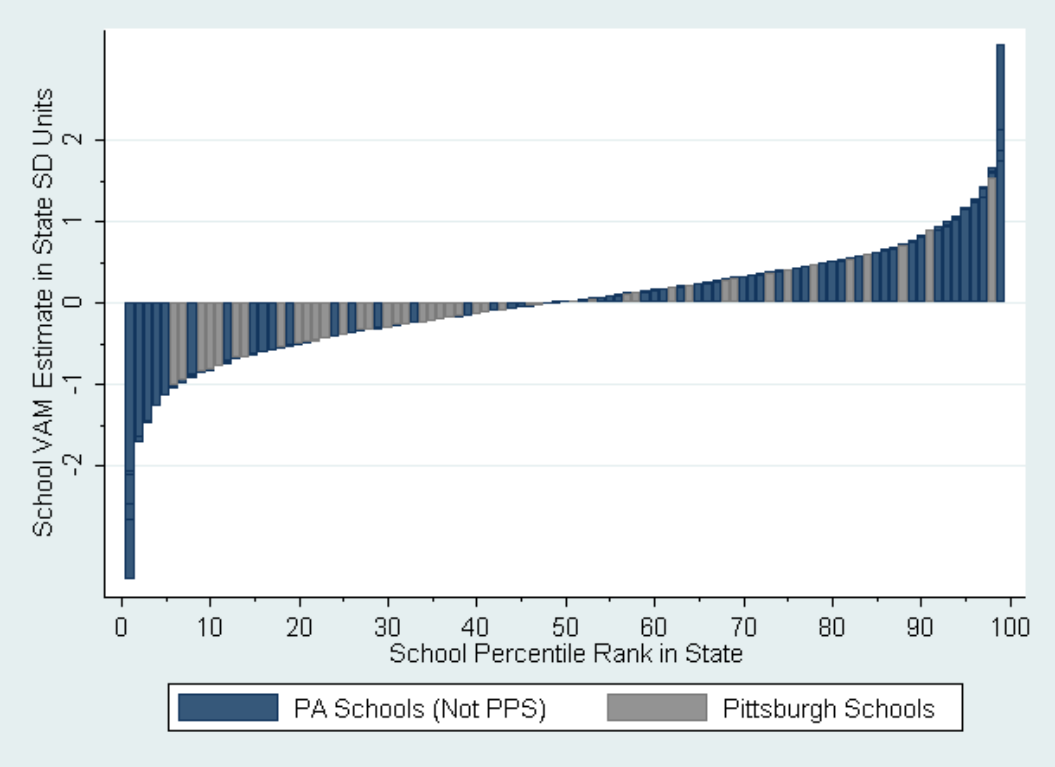
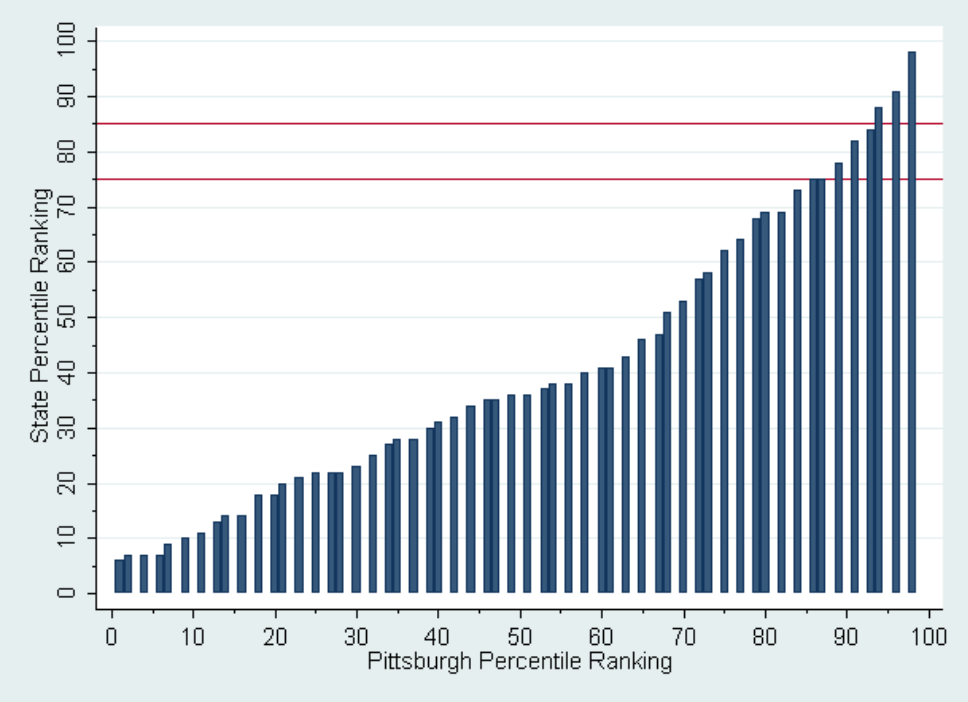


Figure VI.2. State Percentile Composite VAM Ranking of Pittsburgh Schools, 2009-11



Source: Mathematica calculations based on data from the Pennsylvania Department of Education

Note: Each bar represents an individual Pittsburgh school.

As noted earlier, fewer outcome measures are available at the state level than are available in Pittsburgh. In elementary grades, there is no difference in the available outcomes. In middle and high school grades, however, VAM estimates based on assessments not available statewide—CBAs and PSATs—are mapped to the PSSA assessment that is most closely related to them in terms of grade and subject, creating an indirect estimate of statewide performance. Thus, for example, the 7th-grade math CBA is mapped to the statewide distribution on the 7th-grade math PSSA. For high schools, state value-added percentiles are assigned based on a school-level 11th-grade PSSA VAM in the relevant subject, using 8th-grade baseline scores from three years earlier. High-school science is the exception: Because of concern that the 11th-grade PSSA science assessment is not well aligned with PPS science curriculum, PPS has chosen to use the statewide PSSA math value-added result, rather than the statewide science value-added result, as a benchmark for the high-school science CBA VAMs. There is no statewide social studies assessment, so we use Pittsburgh's statewide VAM results on PSSA reading and writing assessments for the purpose of assigning state value-added percentiles for social studies CBA VAMs.

PPS-specific composites that include local assessments alongside PSSAs are mapped to PSSA-based statewide composites as indicated in Table VI.1. In Table VI.1, we show the assessments and other outcomes that are available at the statewide level and how they are used for assigning state value-added percentiles to Pittsburgh composites. Specifically, in each cell we list the component measures of the statewide distribution that are used for each subject and grade level. Altogether, 46 percent of Pittsburgh schools had overall composite effectiveness scores that were statistically different from the median school statewide. These schools included the eight schools placing in the top 25 percent of statewide performance and 18 schools with estimates below median statewide performance.

Table VI.1. The Composition of Statewide Composites, 2009-11

	Elementary School Grades K to 5	Middle School Grades 6 to 8	High School Grades 9 to 12
Test-based Measures			
Math composite	PSSA math (4,5)	PSSA math (6,7,8)	PSSA math (11)
English/language arts composite	PSSA reading (4,5) PSSA writing (5)	PSSA reading (6,7,8) PSSA writing (8)	PSSA reading (11) PSSA writing (11)
Science composite	PSSA science (4)	PSSA science (8)	PSSA math (11)
Social studies composite	n/a	PSSA reading (6,7,8) PSSA writing (8)	PSSA reading (11) PSSA writing (11)

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

C. Comparison to PVAAS

The VAMs described here have similarities and differences with the Pennsylvania Value Added Assessment System (PVAAS). Like the VAMs developed for PPS, PVAAS provides value-added data based on statistical analyses of PSSA scores in grades 4 to 8 and 11. It does so by linking students' achievement records in multiple subjects over time to measure whether cohorts make a year's worth of improvement in core subject areas.

But PVAAS also differs in key respects. First, it calculates value-added for grade-subject combinations in each school, but it does not calculate value-added for teachers. Second, it includes students' entire assessment histories but does not control for socioeconomic or demographic factors. Third, the primary metric of reporting of PVAAS results (its color scheme) uses as its comparative reference the statewide value-added distribution in 2006, rather than the current value-added distribution. PVAAS defines a year of achievement growth as the average value-added statewide in 2006. Fourth, PVAAS examines only one year of teaching at a time, rather than two years (as in our school VAMs) or three years (as in our teacher VAMs).[16]

We compared PVAAS estimates of average gain over tested grades with one-year estimates from our own models, and found results to be highly correlated in nearly all cases; results are in Appendix A.2.

Although correlations with the PVAAS numeric averages are high, the PVAAS results that get the most attention—the color codes that measures performance against the 2006 statewide distribution—tend to be systematically inflated relative to our VAM estimates. This is likely to be the effect of PVAAS' use of a 2006 benchmark. Scores statewide have risen since 2006, so that more than half of schools are above that benchmark. Our VAMs compare PPS schools to current statewide average performance.

[16] PVAAS uses a different methodology for 11th-grade value-added estimates, because state assessments do not yet exist in 10th grade.

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

A. Teacher VAM Results

The summary results for the teacher VAMs are displayed in Table VII.1. The results are displayed by grade and assessment. On average across all assessments, using a 95 percent confidence interval, we can distinguish 33 percent of teachers from the PPS average. The dispersion of value-added estimates varies by grade, subject, and assessment. At the extremes, the 90th percentile teacher raises achievement on the eighth-grade math CBA by 0.68 standard deviations compared to the average teacher, while the 90th-percentile teacher raises achievement on the 11th grade reading PSSA by only 0.06 standard deviations. One possible explanation for the very small variance in value-added on the 11th-grade reading PSSA is that the test may not be well aligned with the 11th-grade English curriculum.

These average effect sizes can be interpreted in terms of the average gains typical students at different grade levels are expected to make from year to year. Figure VII.1 shows how the teacher effects for PSSA outcomes can be described in terms of the approximate proportion of the average amount of learning achieved by a typical student nationally in that grade and subject. These estimates are based on the expected gains by grade and subject reported by Hill et al. (2008), based on seven nationally normed standardized tests. Their accuracy in the PSSA context in Pittsburgh depends on an assumption that the variance of learning of students in Pittsburgh is approximately equivalent to the variance of student learning of the students in the national samples used by Hill et al. to generate their estimates. CBA outcomes are not included in this table because they are specific to Pittsburgh and are likely not comparable to the nationally normed assessments used to estimate gains of typical students.

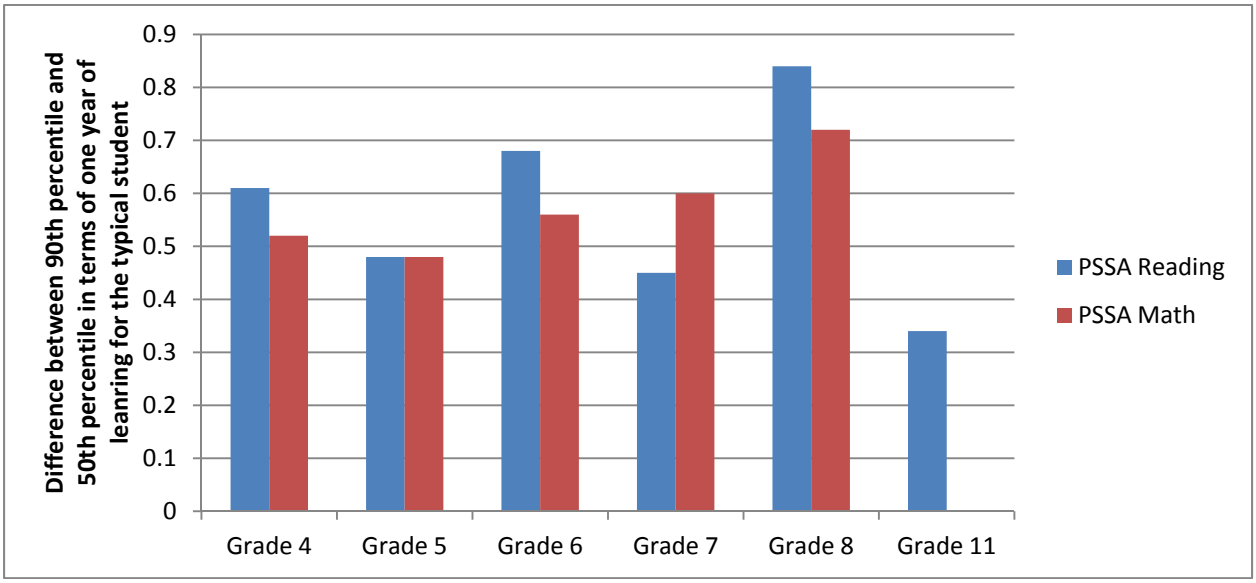
On average across grades and subjects, a typical student with a teacher at the 90th percentile teacher learns approximately an additional 57 percent of a typical year of learning achieved in nationally normed assessments, relative to how much is learned by a student with the median Pittsburgh teacher.

This varies by grade and subject, however, because students make relatively larger gains in some grade levels and subjects than others. Hill et al. (2008) found that annual gains are largest at lower grade levels, with average gains between fourth and fifth grades of 0.40 standard deviations in reading and 0.56 standard deviations in math. In Pittsburgh, the 90th percentile teacher increased fifth grade math achievement by 0.27 standard deviations more than the average teacher, or approximately 48 percent of what a typical fifth grade student would be expected to gain in math during the school year. Meanwhile, the 90th percentile teacher increased eighth grade achievement by 0.23 standard deviations on the math PSSA compared to the average Pittsburgh teacher. This equates to approximately 72 percent of what a typical eighth grade student would be expected to gain in math during the school year.

Table VII.1. Teacher VAM Results, by Outcome 2008-11

Outcome	Grade	Adj. R-squared	Years of teaching	Teachers	Difference between 90th percentile and 50th percentile in Z-score units	SD of teacher effects	Mean standard error	Statistically significant effects (95% CI)
PSSA Math	4	0.72	3	80	0.27	0.21	0.10	0.34
PSSA Reading	4	0.73	3	86	0.22	0.17	0.10	0.24
PSSA Science	4	0.69	3	50	0.31	0.24	0.09	0.46
PSSA Math	5	0.80	3	69	0.27	0.21	0.10	0.32
PSSA Reading	5	0.76	3	84	0.19	0.15	0.10	0.27
PSSA Writing	5	0.51	3	73	0.36	0.28	0.12	0.44
PSSA Math	6	0.76	3	78	0.23	0.18	0.09	0.32
PSSA Reading	6	0.74	3	98	0.22	0.17	0.10	0.31
CBA Math	6	0.60	2	54	0.51	0.40	0.15	0.39
CBA English	6	0.54	3	63	0.26	0.20	0.12	0.21
CBA Earth Science	6	0.58	3	48	0.53	0.41	0.11	0.69
PSSA Math	7	0.80	3	72	0.18	0.14	0.08	0.32
PSSA Reading	7	0.75	3	92	0.10	0.08	0.08	0.10
CBA Math	7	0.56	2	53	0.36	0.28	0.14	0.30
CBA English	7	0.54	3	66	0.18	0.14	0.11	0.14
CBA Life Science	7	0.66	3	37	0.45	0.35	0.09	0.59
PSSA Math	8	0.81	3	62	0.23	0.18	0.08	0.39
PSSA Reading	8	0.75	3	75	0.22	0.17	0.09	0.29
PSSA Science	8	0.75	3	37	0.18	0.14	0.07	0.32
PSSA Writing	8	0.58	3	74	0.29	0.23	0.12	0.27
CBA Math	8	0.40	2	32	0.68	0.53	0.19	0.41
CBA English	8	0.58	3	54	0.33	0.26	0.11	0.35
CBA Physics	8	0.63	2	31	0.59	0.46	0.11	0.48
CBA US History	8	0.66	2	31	0.53	0.41	0.17	0.52
CBA Algebra I/AB-BC	9	0.50	3	50	0.54	0.42	0.16	0.38
CBA ELA I	9	0.51	3	42	0.33	0.26	0.14	0.26
CBA Biology	9	0.57	3	23	0.44	0.34	0.13	0.52
CBA Civics	9	0.57	3	30	0.44	0.34	0.10	0.60
CBA Geometry/AB-BC	10	0.67	2	31	0.36	0.28	0.13	0.32
CBA ELA II	10	0.57	2	27	0.19	0.15	0.13	0.11
CBA Chemistry	10	0.45	2	21	0.35	0.27	0.15	0.24
CBA World History	10	0.49	2	23	0.47	0.37	0.18	0.43
CBA Algebra II	11	0.54	2	32	0.65	0.51	0.21	0.50
CBA ELA III	11	0.54	2	30	0.36	0.28	0.17	0.20
CBA Physics	11	0.42	1	14	0.50	0.39	0.27	0.21
CBA US History	11	0.51	2	19	0.42	0.33	0.15	0.32
PSSA Reading	11	0.68	2	38	0.06	0.05	0.08	0.00
PSSA Writing	11	0.43	2	37	0.19	0.15	0.14	0.08
CBA ELA IV/AA Lit	12	0.47	2	25	0.17	0.13	0.15	0.04

Figure VII.1. Teacher VAM Results for Math and Reading PSSAs Expressed in Fractions of a Year of Learning: Difference between Median Teacher and 90th-percentile Teacher in Pittsburgh



The overall composite effectiveness score for teachers is calculated based on the assessment VAMs described by Table VII.1. Overall composite effectiveness measures are available for 788 teachers, of whom 31 percent can be distinguished statistically from typical performance using a 95 percent confidence level. Note that the shrinkage adjustment means that in practice, even less than five percent of teachers are likely to be falsely identified as significantly above or below average. In the one case in which no teachers can be distinguished from average—11th-grade PSSA reading—a joint test of the significance of results (F-test) confirms that nonetheless the distribution is not random.

B. School VAM Results

1. School VAM Results for Grades 4-5

The results of the school level VAMs for grades 4 and 5 are presented in Table VII.2. These models are based on two years of data. Across all fourth and fifth grade assessments, the 90th percentile school raises achievement by between 0.14 and 0.33 standard deviations compared to the average school. This can be interpreted in terms of the expected gains of a typical student in each grade and subject, based on the estimates reported by Hill et al. (2008). For example, a fourth grade student at the 90th percentile school learns, on average, approximately 37 percent more in a year than a typical year's worth of learning. The gains are greater in fifth grade, with a student at the 90th percentile school learning 52 percent more in math and 58 percent more in reading than a typical year of learning. A test of the joint significance of results for each VAM confirms that results are not merely random distributions.

Table VII.2 School VAM Results for Grades 4 to 5, by Outcome

Outcome	Grade	Adj. R-squared	Students	Schools	Difference between 90th and 50th percentile schools: In Z-score units	Difference between 90th and 50th percentile schools: In Terms of One Year of Learning for a Typical Student	Mean standard error	Statistically significant effects (95% CI)
PSSA Math	4	0.71	3463	38	0.19	0.37	0.07	0.34
PSSA Reading	4	0.72	3460	38	0.14	0.39	0.06	0.29
PSSA Science	4	0.71	3444	38	0.26	NA	0.07	0.45
PSSA Math	5	0.79	3409	38	0.29	0.52	0.06	0.47
PSSA Reading	5	0.75	3394	38	0.23	0.58	0.07	0.55
PSSA Writing	5	0.49	3362	38	0.33	NA	0.10	0.50

Note: Difference between 90th percentile and 50th percentile school in terms of one year of learning is based on estimates from Hill et al. (2008), as described in the text above. The estimates of a typical student's gains in one year, based on seven nationally normed standardized tests, are reported in Table VII.2.

2. School VAM Results for Grades 6-8

The results of the school level VAMs for grades 6 through 8 are presented in Table VII.3. These models are also based on two years of data. On average, 36 percent of schools can be distinguished from average. A student at the 90th percentile school learns, on average, between 44 and 60 percent more than the amount learned in a typical school.

There is more variation in school effects across assessments in middle school than in elementary. For example, the 90th percentile school raises achievement on the eighth grade US History CBA by 0.50 standard deviations compared to average, while the 90th-percentile school on the seventh-grade math CBA improves results by only 0.06 standard deviations relative to a typical school.

3. School VAM Results for Grades 9-12

The results of the school level VAMs for grades 9 through 12 are presented in Table VII.4. These models are also based on two years of data (with the exception of the 11th grade physics CBA model, for which only one year of data is available). On average, 34 percent of schools can be distinguished from average. There are several assessments for which no schools could be distinguished from average, including the eleventh grade ELA CBA and reading PSAT, but this is at least partly due to the fact that Pittsburgh has only a small number of schools serving grades 9-12.

Table VII.3 School VAM Results for Grades 6 to 8, by Outcome

Outcome	Grade	Adj. R-squared	Students	Schools	Difference between 90th and 50th percentile schools: In Z-score units	Difference between 90th and 50th percentile schools: In Terms of One Year of Learning for a Typical Student	Mean standard error	Statistically significant effects (95% CI)
PSSA Math	6	0.76	3239	28	0.18	0.44	0.06	0.36
PSSA Reading	6	0.74	3222	28	0.14	0.44	0.06	0.25
CBA Math	6	0.55	2449	27	0.21	NA	0.10	0.22
CBA English	6	0.56	2743	27	0.35	NA	0.10	0.37
CBA Earth Science	6	0.60	2947	27	0.47	NA	0.09	0.52
PSSA Math	7	0.79	3288	27	0.18	0.60	0.06	0.41
PSSA Reading	7	0.74	3282	27	0.08	0.35	0.05	0.04
CBA Math	7	0.55	2727	27	0.06	NA	0.07	0.00
CBA English	7	0.53	2718	26	0.24	NA	0.09	0.27
CBA Life Science	7	0.66	3038	27	0.40	NA	0.08	0.63
PSSA Math	8	0.80	3252	27	0.18	0.56	0.05	0.48
PSSA Reading	8	0.74	3260	27	0.15	0.58	0.06	0.33
PSSA Science	8	0.75	3227	27	0.12	NA	0.06	0.19
PSSA Writing	8	0.57	3218	27	0.28	NA	0.08	0.44
CBA Math	8	0.37	1804	24	0.47	NA	0.14	0.46
CBA English	8	0.57	2862	27	0.14	NA	0.08	0.11
CBA Physics	8	0.60	2978	26	0.46	NA	0.09	0.62
CBA US History	8	0.64	2855	25	0.50	NA	0.09	0.76

Note: Difference between 90th percentile and 50th percentile school in terms of one year of learning is based on estimates from Hill et al. (2008), as described in the text above. The estimates of a typical student's gains in one year, based on seven nationally normed standardized tests, are reported in Table VII.2.

Conversions to years of learning are largely unavailable in the high school grades, because the great majority of student assessments are locally-developed CBAs. Even the 11th-grade PSSAs must rely on CBAs for baseline controls. We therefore report effect sizes in years of learning only for the 11th-grade PSAT, which uses the 10th-grade PSAT as its baseline. The 90th-percentile school adds an estimated 0.24 years of learning as measured by the 11th-grade PSAT math and 0.16 years of learning as measured by the 11th-grade PSAT reading (as compared to the median school).

Table VII.4. School VAM Results for Grades 9 to 12, by Outcome

Outcome	Grade	Adj. R-squared	Students	Schools	Difference between 90th and 50th percentile schools: In Z-score units	Difference between 90th and 50th percentile schools: In Terms of One Year of Learning for a Typical Student	Mean standard error	Statistically significant effects (95% CI)
CBA Algebra I/AB-BC	9	0.37	1851	12	0.37	NA	0.10	0.67
CBA ELA I	9	0.46	1998	11	0.31	NA	0.08	0.55
CBA Biology	9	0.49	2247	11	0.35	NA	0.08	0.45
CBA Civics	9	0.58	2284	12	0.36	NA	0.07	0.58
CBA Geometry/AB-BC	10	0.62	1849	11	0.37	NA	0.08	0.91
CBA ELA II	10	0.56	1510	10	0.22	NA	0.09	0.40
CBA Chemistry	10	0.41	1352	9	0.31	NA	0.11	0.11
CBA World History	10	0.45	1379	10	0.33	NA	0.11	0.50
PSAT Math	10	0.61	2233	12	0.13	NA	0.06	0.25
PSAT Reading	10	0.63	2234	12	0.08	NA	0.05	0.17
PSAT Writing	10	0.58	2187	12	0.17	NA	0.07	0.25
CBA Algebra II	11	0.46	1427	10	0.45	NA	0.11	0.60
CBA ELA III	11	0.52	1322	10	0.09	NA	0.08	0.00
CBA Physics*	11	0.37	407	8	0.27	NA	0.19	0.13
CBA US History	11	0.43	1174	9	0.28	NA	0.11	0.33
PSSA Math	11	0.71	1466	11	0.21	NA	0.07	0.45
PSSA Reading	11	0.67	1611	11	0.08	NA	0.06	0.09
PSSA Writing	11	0.42	1575	11	0.18	NA	0.09	0.18
PSAT Math	11	0.76	2046	10	0.06	0.24	0.05	0.20
PSAT Reading	11	0.76	2048	10	0.03	0.16	0.04	0.00

*All estimates are based on two years of performance (2009-10 and 2010-11), except 11th grade CBA Physics, which can be estimated only for one year, 2010-11.

Note: Difference between 90th percentile and 50th percentile school in terms of one year of learning is based on estimates from Hill et al. (2008), as described in the text above. The estimates of a typical student's gains in one year, based on seven nationally normed standardized tests, are reported in Table VII.2. Because PSAT assessments are taken at the beginning of the school year, teacher effects are converted to years of learning based on the prior year's expected learning gains. For example, the grade 10 PSAT is converted to years of learning based on what typical students would be expected to learning during ninth grade.

The range of school effects varies in high school as it does in lower grades, with the 90th percentile school raising achievement on the eleventh grade Algebra II CBA by 0.45 standard deviations compared to average and only by 0.03 standard deviations on the eleventh grade reading PSAT.

4. School VAM Results for Non-Test Outcomes

In addition to the value-added models described above, we also ran school level VAMs on two non-test outcomes: attendance and core course pass rate. The results of these models are presented in Table VII.5. These models are based on two years of data. On average, the 90th percentile school raises attendance by between 0.05 and 0.21 standard deviations depending on grade level, with larger effects in the upper grades. Core course pass rates, which were only measured in high schools,

increased by 0.12 standard deviations more in the 90th percentile high school than in the average school.

Table VII.5 School VAM Results, Non-test Outcomes

Outcome	Grades	Students	Schools	Difference between 90th and 50th percentile in Z-score units	Mean standard error	Statistically significant effects (95% CI)
Attendance	1-3	8344	37	0.05	0.04	0.05
Attendance	4-8	16526	48	0.10	0.04	0.33
Attendance	9-12	10645	12	0.21	0.03	0.58
Core Course Pass Rate	9-12	10500	12	0.12	0.03	0.75

The results of the school level test-based composite VAMs are presented by school type and subject in Table VII.6. All types of schools have math and reading composites. Only schools with middle or high school grades have science VAMs, and only schools with high school grades have social studies VAMs. The overall composite includes all test-based assessments but not attendance or core course pass rate. As might be expected, the composite VAMs generally have more power to distinguish schools from average. Schools are more-often distinguished on the composite VAM for math than on the composite VAM for reading.

Table VII.6 Test-Based Composite School VAM Results

		Statistically significant effects (95% CI)				
School Type	Schools	Math	Reading	Science	Social Studies	Overall Composite
K-5	22	0.91	0.45	N/A	N/A	0.55
K-8	19	0.89	0.58	0.95	N/A	0.79
6-8	7	1.00	0.57	1.00	N/A	0.86
9-12	7	0.86	0.57	1.00	0.29	0.57
6-12	4	1.00	1.00	0.50	0.50	0.75

THIS PAGE LEFT BLANK FOR DOUBLE-SIDED PRINTING

VIII. APPLICATIONS TO REWARDS AND RECOGNITION OPPORTUNITIES

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

A. Promise-Readiness Corps Details

The Promise-Readiness Corps is designed to prepare Pittsburgh students to benefit from The Pittsburgh Promise, a scholarship program to help students plan, prepare and pay for education beyond high school (details are available at www.pittsburghpromise.org). Entering 9th graders at most PPS high schools are divided into Promise-Readiness Corps groups, each of which is assigned to a team of teachers who are collectively responsible for the students for two years (9th and 10th grades). Promise-Readiness Corps teams that perform well in value-added terms are eligible for team-wide cash awards. The VAMs used to estimate the effectiveness of the Promise-Readiness Corps teams at improving outcomes for 9th and 10th grade students are similar to the school level VAMs. The same 9th and 10th grade assessments in Table II.2 are used, as well as the non-test based outcomes for attendance and core course pass rate.

The Promise-Readiness Corps VAMs and the school level VAMs differ along three dimensions:

- **Dosage Variables.** The dosage variables apply to the fraction of a school year that a student spends with a Promise-Readiness Corps team rather than enrolled in a school.
- **Class Size and Advanced Courses.** The Promise-Readiness Corps VAMs have measures for class size and for students taking advanced courses as additional control variables, because these variables are presumed to be outside the control of Promise-Readiness Corps teams. In this respect the Promise-Readiness Corps VAMs resemble the teacher VAMs rather than the school VAMs.
- **Comparison Group.** The comparison group used to evaluate the performance of the Promise-Readiness Corps teams is the historical performance of similar PPS students. The comparison group for Promise-Readiness Corps teams and the use of historical data are described in detail below.

The aim of the Promise-Readiness Corps is to improve on the prior performance of the district's own high schools, rather than to outperform other teams or other schools in the state. The performance of Promise-Readiness Corps teams is therefore compared to historical performance of students in PPS from 2007-2010 (i.e., the district's average performance for those grade levels for the three years immediately preceding the creation of the Promise-Readiness Corps). The use of historical student outcomes data allows us to determine whether Promise-Readiness Corps teams are improving student achievement in absolute terms, and prevents the competition for the Promise-Readiness Corps awards from becoming a horse race where improvements in the performance of one Promise-Readiness Corps team negatively affect the performance of other Promise-Readiness Corps teams.

The first Promise-Readiness Corps cohort is evaluated using 2010-11 data on the performance of 9th grade students only, because this cohort has been in existence for only one year. Here we describe the general approach that will be used with two-year cohorts in the future; the first awards, to be made in spring 2012 based on 2010-11 student results, will use a variant of this approach that examines 9th-grade results only (rather than results over two years for students progressing through 9th and 10th grades).

To determine whether the Promise-Readiness Corps teams have on average improved outcomes for their students we need to analyze student scores on a set of assessments that have been given in each of the last several years *and are scaled consistently from year to year*. CBA scales are not consistent from year to year, and 9th and 10th graders do not take any PSSAs. Fortunately, students take the PSAT in the fall of 10th grade and the fall of 11th grade (i.e., shortly after completing 9th and 10th grades), and PSAT scales are highly consistent from year to year. We therefore compare the PSAT scores of students taught by Promise-Readiness Corps teams to the average performance of similar students from 2007-08 to 2009-10 to determine whether Promise-Readiness Corps teams *on average* improved outcomes. This does *not* mean that the PSAT is the sole assessment determining a Promise-Readiness Corps team's value-added. Instead, the PSAT is used to benchmark district-wide performance of Promise-Readiness Corps teams in much the same way that statewide PSSA VAM estimates are used to benchmark district-wide performance of PPS schools. All of the 9th- and 10th-grade VAMs (for CBAs, attendance, and core-course completion) contribute to Promise-Readiness Corps VAM estimates.

The PSAT scores are used in conjunction with the value-added models as follows: We begin by estimating the district-wide change in value-added on the PSAT, comparing the 9th- and 10th-grade cohorts of 2007-08, 2008-09, and 2009-10 (averaged) with the scores of PRC students (who take the PSAT in the fall after 9th grade and in the fall after 10th grade).[17] The change in district-wide PSAT value-added is used to shift the value-added distribution of the Promise-Readiness Corps teams across the full range of CBAs. In other words, the within-district CBA VAM estimates for each Promise-Readiness Corps team determine that team's relative position among Promise-Readiness Corps teams, but the ultimate team VAM result ("coefficient E" in the Promise-Readiness Corps formula) can shift based on district-wide PSAT VAM performance. A district-wide improvement in PSAT value-added will shift upward the CBA-based VAM estimates of all of the Promise-Readiness Corps teams. If Promise-Readiness Corps students score very high on the PSAT as compared to past students, it will be theoretically possible for all teams to be ranked as above average in value-added, and for all of them to receive Promise-Readiness Corps results that entail financial awards.

The non-test based outcomes used to evaluate Promise-Readiness Corps teams are similarly anchored using the distribution over the three prior years of the non-test based outcomes. It is assumed that PPS is ensuring that standards for attendance and passing of core classes are the same from year to year. (If attendance rates or passing rates were artificially inflated as compared with past rates, the number of awards could increase without a true increase in performance.)

[17] We use the 8th grade PSSA scores to adjust for the prior ability of the students taking the PSAT. To account for the possibility that the scoring or scale of the PSSA may change over time, we standardize the PSSA scores based on the statewide distribution before making the adjustment.

Once the value-added distributions of the test and non-test based outcomes are modified based on the performance relative to historical standards, the value-added performance is combined using weights specified by PPS into one composite measure, which is used to inform award size (after being converted to "coefficient E"). The current weighting, determined by a working group of teachers and district staff, assigns 50% of the Promise-Readiness Corps composite to be based on non-test-based VAM results and 50% to be based on test-based VAM results, with precision weighting used to determine the relative weights among individual VAMs within the test-based and non-test-based categories. PRC team members will receive VAM reports on their team's results prior to the distribution of awards.

B. Students and Teachers Achieving Results (STAR) Details

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 will receive awards applauding their achievement. It is the intention of the STAR schools plan to recognize at least eight schools a year. Accordingly, if fewer than eight PPS schools place in the top 15 percent, the next highest-ranked schools up to that number will be identified in order of student growth, as long as these schools place in the top 25 percent of growth in the State. The first STAR schools will be named in the 2012-13 school year, based on achievement results concluding with spring 2012.

Pittsburgh's collective bargaining agreement requires a statewide comparison for determination of STAR awards, so STAR VAMs will include only outcomes that are available on a statewide basis. Because STAR requires that only statewide assessments will be used, this means that VAMs will be estimated using only PSSA scores and (for high schools) data on holding power. The VAMs for STAR thus differ from those used in the regular school VAM reporting, because the statewide data will be the only source of information. Specifically, we will estimate statewide VAMs and develop for each grade range (4-5, 6-8, and 9-12) a single composite measure including all of the statewide VAMs. Schools with grade configurations of K-8 or 6-12 receive composite scores that include all STAR outcomes from the relevant grades. We will then use the composite VAMs to determine which schools place in the top 15 (or 25) percent of the statewide distribution.

In Table VIII.1, we show which assessments constitute the basis for identifying STAR schools in each grade range. Note that, per the decision of PPS, this includes the 11th-grade PSSA science assessment, which is not included in other VAM analyses. STAR awards will be determined based on the overall composite, combining test-score VAMs and holding power VAM.

Table VIII.1. Assessments Used to Determine the STAR Award System by Grade Range, 2011-12

	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 reading (4,5) PSSA writing (5) PSSA science (4)	PSSA math (6,7,8) PSSA reading (6,7,8) PSSA writing (8) PSSA science (8)	PSSA math (11) PSSA reading (11) PSSA writing (11) PSSA science (11) Holding power (9-11)

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

The statewide VAMs will be the same as those described in Chapter VI. Specifically, they will include the same baseline/background variables and include two years of student performance whenever possible. The statewide VAMs for STAR also will be combined into composite measures using the same approach that is described in Chapter V, consisting of an initial standardization of the individual VAM distributions followed by the application of PPS-determined weights that have been informed by precision calculations and teacher and administrator input.

REFERENCES

Goldhaber, D., and D. Chaplin. "Assessing the 'Rothstein Test.' Does it Really Show Teacher Value-added Models are Biased?" CALDER Working Paper 71. 2012.

Hanushek, E. A., and S. G. Rivkin. "Do Disadvantaged Urban Schools Lose their Best Teachers?" The Urban Institute, National Center for Analysis of Longitudinal Data in Education Research, 2008.

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

Isenberg, E. and H. Hock. "Measuring School and Teacher Value Added for IMPACT and TEAM in DC Public Schools." Final report to DC Public Education Fund, District of Columbia Public Schools, and New Leaders for New Schools. Washington, DC: Mathematica Policy Research, 2010.

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

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

McCaffrey, D. F., Sass, T. R., Lockwood, J. R., 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 American Statistical Association*, vol. 78, no. 381, 1983, pp. 47-55.

Potamites, L., Booker, K., Chaplin, D., and Isenberg, E. "Measuring School and Teacher Effectiveness in the EPIC Charter School Consortium – Year 2." Final report submitted to New Leaders for New Schools. Washington, DC: Mathematica Policy Research, 2009.

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

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

Schochet, P. Z. and Chiang, H. S. *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.

THIS PAGE LEFT BLANK FOR DOUBLE-SIDED PRINTING

APPENDIX TABLES

Table A.1. Precision of Teacher VAM Results for Grades 6 to 8, by Years of Teaching Included

Outcome	Grades	Mean Standard Error	Statistically Significant Effects
One-year model (2009-10)			
PSSA math	6-8	0.07	0.38
PSSA reading	6-8	0.08	0.13
PSSA science	8	0.08	0.10
PSSA writing	8	0.13	0.34
Three-year model (2007-10)			
PSSA math	6-8	0.04	0.52
PSSA reading	6-8	0.05	0.36
PSSA science	8	0.05	0.56
PSSA writing	8	0.08	0.73

Table A.2. Correlation of Mathematica and PVAAS VAM Estimates, Grades 4 to 8

		Correlation to PVAAS	
	Grades Included	2009-10	2008-09
Math	4-8	**0.62**	**0.58**
Reading	4-8	**0.81**	**0.56**
Science	4	**0.92**	**0.92**
Science	8	0.17	**0.79**
Writing	5	**0.94**	**0.68**
Writing	8	**0.78**	**0.82**

Note: PVAAS statistic is the mean NCE gain over grades relative to the state (math and reading) or the school effect (science and writing). Because the mean NCE gain over grades includes all grades between 4 and 8 offered at a school, we create a weighted average of VAM scores for K-8 schools. The weight is 0.4 for the grade 4-5 score and 0.6 for the grade 6-8 score. Bold indicates statistical significance using a 95 percent confidence interval.

MATHEMATICA
Policy Research

www.mathematica-mpr.com

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

Princeton, NJ ■ Ann Arbor, MI ■ Cambridge, MA ■ Chicago, IL ■ Oakland, CA ■ Washington, DC