Estimating Teacher and School Effectiveness in Pittsburgh: Value- Added Modeling and Results

Final Report

November 15, 2010

Stephen Lipscomb
Brian Gill
Kevin Booker
Matthew Johnson

MATHEMATICA
Policy Research, Inc.

Estimating Teacher and School Effectiveness in Pittsburgh: Value-Added Modeling and Results

Final Report

November 15, 2010

Stephen Lipscomb
Brian Gill
Kevin Booker
Matthew Johnson

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

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

MATHEMATICA
Policy Research, Inc.

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. Although we cannot name everyone, we want to especially thank Mary Wolfson, Paulette Poncelet, Grace Tan, Sam Franklin, Jack Garrow, and Bill Hileman for their support for the work. We are grateful to Joe Charnock for providing us with the data used in this report. Howard Nelson, Angela Minnici, and Rob Weil from the American Federation of Teachers, along with John Tyler and Drew Gitomer, also provided useful input to an earlier version of this report.

Several staff at Mathematica Policy Research contributed to this report as well. We thank Duncan Chaplin, Hanley Chiang, John Deke, and Eric Isenberg for their suggestions and Rosie Malsberger for organizing the data to set up the model estimation.

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 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 95th 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

SUMMARY

At the request of Pittsburgh Public Schools (PPS) and the Pittsburgh Federation of Teachers (PFT), Mathematica is developing 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. The analyses described in this report are intended as an early step in a multi-year project that aims to produce fair, valid, reliable, and robust estimates of the contributions of individual teachers and schools in Pittsburgh to the achievement growth of their students, ultimately including as many grades, subjects, schools, and teachers as possible.

Data included in this report: student outcomes and teachers

This report substantially expands the range of student outcomes incorporated in VAMs, beyond those examined in our earlier report (Lipscomb, Gill, and Booker, 2010[1]). In addition to scores on the Pennsylvania System of School Assessment (PSSA), this report also applies VAMs to Pittsburgh's own curriculum-based assessments (CBAs) in various subjects and courses in grades 6-10; to PSAT assessments conducted in grades 10 and 11; and to Scholastic Reading Inventory (SRI) assessments in grades 6 to 10. We also create composite value-added measures that incorporate results for each of the assessment outcomes relevant to each school. In addition, school-level VAMs are estimated for the non-test outcomes of student attendance and credit accumulation.

This report describes the methods for measuring teacher and school effectiveness using those student outcomes and the findings from our analyses to date. The report does not identify any individual schools or teachers as yet, instead describing in general terms the distribution of value-added estimates for teachers and schools across the district.

The findings in this report reinforce the findings of the August report (Lipscomb, Gill, and Booker, 2010) that *VAM estimates can provide meaningful information about teacher and school performance in Pittsburgh*—and they confirm that *VAMs can be usefully applied not only to state accountability tests and nationally normed assessments, but also to locally developed assessments and to non-test measures of student outcomes such as attendance and credit accumulation.*

We do not find evidence suggesting that any of the locally developed CBAs must be excluded from incorporation in VAMs. The addition of results from CBAs, SRI, PSAT, and science and writing PSSA assessments have allowed far more teachers to be included in value-added analyses, and have dramatically broadened the range of outcomes incorporated in school-level VAM estimates (which now include attendance and credit accumulation as well as many more assessment outcomes). Our estimates for the 2009-10 year apply to 646 teachers, an increase of more than 40 percent from our August report. In future years, there should be opportunities to increase this number further, because in the historical data there are some CBAs that are entirely missing for a

[1] Lipscomb, Gill, and Booker (2010) is a draft report that was intended for the internal uses of Pittsburgh Public Schools.

small number of schools. In the future, the district should ensure that CBA results are consistently collected from all schools.

Control variables included in VAMs

The VAMs used in this report include statistical controls for each student's previous-year assessments. They also include controls for other student characteristics that are included in PPS records and might affect achievement growth, outside the influence of individual teachers. These include special education status, gifted status, poverty status (as indicated by eligibility for free- or reduced-price lunch), English language learner status, gender, mobility, prior year attendance and disciplinary incidents, date of birth, and race/ethnicity. A notable refinement since our August report is the addition of codes for various specific types of disabilities, rather than a single code for any student with a disability.

PPS and PFT asked us to examine the extent to which the inclusion or exclusion of a variable for race/ethnicity affects the value-added estimates. We find that the inclusion of a statistical control for a student's race/ethnicity makes a measurable difference in the results of many of the VAMs (most prominently in high school). Standard practice in social science research is to include race in the analysis because it serves as a proxy for disadvantages that African-American students experience outside of school (e.g., access to family and neighborhood resources). Omitting the variable for race/ethnicity therefore could produce value-added estimates that are biased against teachers who serve higher-than-average proportions of African-American students. Given this finding, we recommend continuing to include race/ethnicity as one of the control variables in the VAMs.

We are continuing to examine the possibility of including additional school-level control variables in the VAMs. This is not a straightforward task, for technical reasons that are discussed in Chapter III. We expect to have further information on including school-level characteristics as controls prior to the release of VAM results next spring.

Precision of VAM estimates

Consistent with findings on teacher quality in the existing research literature, our findings show substantial differences in teacher effectiveness across the district, as measured by value-added in a wide range of grades and subjects. Pittsburgh's most-effective teachers are producing gains in student achievement that are large enough that, if accumulated over several years without decay (admittedly a strong assumption), could erase achievement gaps between black students and white students, or between Pittsburgh students and statewide averages (Lipscomb, Gill, and Booker, 2010). We also find substantial variation in the value-added of individual schools across the district—although there is more variation among teachers than among schools.

Value-added estimates, like any measure of teacher or student performance, are subject to some amount of random variation (Schochet and Chiang, 2010). This variation comes from idiosyncratic differences in the unmeasured abilities of the students in a teacher's class or at a school from one year to the next. If random variation is large, VAM estimates will not reliably identify high and low-performing teachers or schools.

Table S.1 shows the proportion of schools and teachers whose effects can be distinguished from average (either above or below) with 95 percent confidence, when using one, two, or three years of teaching data and a subset of the student assessments examined. (In this report, unlike the August report, we have employed 95 percent confidence intervals in nearly all cases, to promote greater caution in interpretation of results. PPS and PFT may wish to consider the use of less stringent confidence intervals for some purposes in future analyses.)

Table S.1 Share of School and Teacher Effects that are Distinguishable from Average

			School Effects		Teacher Effects	
Grades	Years of Teaching	Assessments	Math	Reading	Math	Reading/ English
4-5	One year	PSSA	26%	18%	35%	24%
4-5	Two years		41%	23%	n/a	n/a
4-5	Three years		n/a	n/a	48%	29%
6-8	One year	PSSA	65%	48%	38%	13%
6-8	Two years		73%	50%	n/a	n/a
6-8	Three years		n/a	n/a	52%	36%
9-10	One year	CBA Geometry and English II	4 of 9	0 of 9	41%	19%
9-10	Two years		n/a	n/a	n/a	n/a
9-10	Three years		n/a	n/a	n/a	n/a

Note: Findings are based on a 95 percent confidence interval.

Comparable data on the other student outcomes is presented in Chapters III (on teachers) and IV (on schools). For most (but not all) outcomes examined in this report, VAMs based on a single year of teaching reliably distinguish the highest- and lowest-performing teachers and schools from average, suggesting that the one-year models are diagnostic for some purposes. But the reliability of teacher-level value-added estimates improves substantially when examining three years of teaching rather than one year. In some cases, using multiple years of data also leads to increases in the proportion of schools that can be reliably distinguished from average performance across the district. We therefore recommend using VAM estimates that are based on two or three years of teaching for high-stakes decisions.

Year-to-year estimates of each teacher's value added are correlated at levels that are similar to those found in other studies outside of Pittsburgh. With a few notable exceptions, schools with high value-added estimates in one year tend to have high value-added estimates in the next year, as shown in Table S.2. VAM estimates for teachers are likewise positively correlated from year to year (results shown in Chapter III).

Table S.2. Year-to-Year Correlations of School-Level VAM Estimates

Subject	Elementary Grades (PSSA)	Middle Grades (PSSA)	High-School Grades (9th-grade CBA)
Math	0.44	0.86	0.43
Reading/English	0.11	0.76	0.86
Science	0.36	0.13	0.71
Writing	0.57	0.50	n/a

Comparison to PVAAS

School-level value-added estimates for PSSA assessments using our models generally correlate reasonably well with value-added estimates produced by the Pennsylvania Value-Added Assessment System (PVAAS) using statewide data. In 2009-10, our VAM estimates for grades 4-8 showed correlations to PVAAS estimates ranging from 0.62 to 0.94, with the exception of 8th-grade science, where the correlation was only 0.17. A primary concern about PVAAS estimates, as noted in our previous report, is the low within-school correlation of estimates from one year to the next (in math and reading).

School value-added by race

Individual teachers do not serve enough students to allow useful VAM estimates for subsets of student populations, but schools are large enough that it may be possible to examine whether a school's value-added varies for different kinds of students (if the relevant groups of students are large enough). At the request of PPS, we examined the extent to which schools showed significantly different value-added estimates for their white students and African-American students. We found that for the great majority of Pittsburgh schools, composite value-added estimates did not differ by statistically significant margins for students of different races. In grades 4-5, only two of 28 schools had value-added estimates that differed significantly by race; this number could have shown differences by chance alone. In grades 6-8, five of 22 schools showed value-added estimates that differed significantly by race. Notably, the differences by race were not always in the same direction: some schools showed higher value-added for African-American students, while others showed higher value-added for white students.

High-school VAMs and student attrition

High-school VAMs pose special challenges, because they can be affected by the inability to include students who drop out prior to the administration of a test. Our previous analysis of 11th-grade PSSA scores (using 8th-grade baselines) suggested that VAM estimates of these outcomes could produce biased results—artificially favoring schools with higher dropout rates—unless adjustments are made to address differential attrition (Lipscomb, Gill, and Booker, 2010).

New analyses are more encouraging: While we do observe a positive correlation between dropout rates and high school VAM estimates for some 2009-10 outcome measures, the pattern is not consistent, and for many outcome measures the correlation is negative. Therefore, at this point we do not find evidence that differential attrition is substantially biasing the high school VAM estimates.

High-school VAMs and the PRC bonus calculation

The analyses of 9th- and 10th-grade outcomes suggest that VAMs using all of the outcomes could potentially be included in a PRC bonus formula, with the caveat that the district will need to ensure that standards for measuring attendance and awarding course credits are consistent across schools. Although most or all of the outcomes can be included in VAMs if desired, we do not believe that their scales are sufficiently consistent from year to year to allow a VAM analysis that uses a historical baseline as the standard for performance.

A notable exception to this rule is the PSAT, which is designed to have consistent scales over time. Indeed, district-wide changes in PSAT scores might be used as a way of assessing the district's overall improvement in value-added. These district-wide changes might then be used to determine the total size of the bonus pool (and/or the number of teams eligible to win bonuses) available to Promise Readiness Corps (PRC) teams, for example. This would allow the district to avoid imposing a "zero-sum game" in the awarding of bonuses: If performance across the district improved on the PSAT, then bonuses would potentially be available to most or even all of the PRC teams.

In future years, the PRC bonus formula might become less dependent on PSAT scores, because the implementation of statewide Keystone exams permits the calculation of value-added estimates for PRC teams relative to the performance of the state as a whole. The performance of the rest of the state would then become the benchmark, rather than the district's own performance on the PSAT in prior years. This, of course, is a policy decision for PPS and PFT to consider.

Next steps

We will continue to refine the VAMs to improve precision and reduce bias. The state of the art in value-added modeling changes rapidly, and we intend to ensure that Pittsburgh's VAMs are updated with refinements developed by our own colleagues and others in the field. Some of these potential refinements are already on the agenda for upcoming work. Specific challenges that we plan to address include the following:

By November 30, 2010:

- We will provide an Excel spreadsheet with identified VAM results for teachers who are potentially eligible for selection into career-ladder roles..

By March 2011:

- We will examine the range in the number of students across schools who are excluded from CBA analyses because they are taking advanced versions of the relevant courses, and we will assess the possible implications of those exclusions.
- We will complete the development a composite value-added measure for teachers that includes information from multiple student outcomes for teachers for whom multiple outcomes are available.
- If the Pennsylvania Department of Education (PDE) provides statewide, longitudinal, student-level data by December 1, we will run PSSA-based VAMs using statewide rather than within-district data, and report those results by March.
- We will assist PPS and PFT staff in the development of a formula to determine the district-wide bonus pool for PRC cohorts. The formula may rely on changes over time in district-wide PSAT results, which are likely to have a scale that is consistent across years.
- We will refine VAM reporting templates for teachers and schools and develop an information booklet that explains VAM concepts in non-technical terms.
- We will examine whether the kindergarten DIBELS assessments can be used to validate a retrospective elementary school VAM that is based on gains from grades 3 to 5. If so, it would increase confidence that VAM estimates using only a subset of elementary

grades nevertheless provide a reasonable view of performance across grades K to 5. The exploratory examination of gains achieved between kindergarten DIBELS assessments and third-grade PSSA scores may also open the possibility of extending the elementary-school VAMs to incorporate an analysis of third-grade results that serves as a proxy for gains from kindergarten. The plans for this analysis are described in Chapter IV.

During 2011:

- In April 2011 we will participate in the event that will communicate school- and teacher-level VAM results to staff across PPS.
- We will examine associations of VAM results to several other measures, including student perceptions data from pilot surveys (if available from Cambridge Education), pilot data from teacher working conditions surveys, and RISE classroom performance evaluations. Proposed plans for these exploratory analyses will be presented by RAA to the EET Steering Committee on December 10, 2010.

In conclusion, we look forward to working with PPS and the PFT in continuing to develop the value-added models to their fullest potential.

I. VALUE- ADDED MODELS FOR THE PITTSBURGH PUBLIC SCHOOLS

A value-added model (VAM) aims to measure the contribution of a teacher or school to student achievement growth. It 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 method is similar conceptually to measuring the average test score gains of students taught by a teacher or school and then taking steps to address additional factors that are outside the control of a teacher or school. For example, if students with learning disabilities tend to grow less on average than other students, a teacher with a disproportionate number of these students may make meaningful contributions toward their achievement growth even if these students score below average on the end-of-year test. Value-added models address these concerns by accounting for other factors that affect student achievement gains. The resulting estimates provide information on teacher or school contributions to student learning that can be used for a variety of purposes, such as helping a school district better target professional development, select teachers for career-ladder positions, and improve teacher evaluation and compensation systems. Our work in estimating teacher and school value-added in Pittsburgh supports the larger, joint efforts of Pittsburgh Public Schools (PPS) and the Pittsburgh Federation of Teachers (PFT) to "empower effective teachers" through applications like these.

This report describes VAMs for teachers and schools in Pittsburgh and extends the analysis of the previous report (Lipscomb, Gill, and Booker, August 2010) by (1) updating our previous results with the most recent year of data; (2) expanding the number of tests used in our VAM; and (3) applying VAMs using student attendance and course completion as outcome measures for schools. We incorporated student outcomes for Pennsylvania System of School Assessment (PSSA) assessments in reading and math from the most recent school year (2009-10) and created value-added measures based on other tests. These include PSSA assessments in science and writing, PSAT scores measuring college readiness, Scholastic Reading Inventory (SRI) computer-administered reading assessments, and curriculum-based assessments (CBAs)—Pittsburgh's home-grown exams for subjects and courses offered in grades 6 through 10. This expansion of analysis allows the calculation of value-added estimates for a much larger number of Pittsburgh's teachers. Many of Pittsburgh's teachers now have more than one value-added estimate because their students have more than one relevant assessment that can be included in a VAM.

We expand the value-added methodology for school performance by using student attendance and course completion as outcome measures. This involves an extension beyond the typical application of value-added models to student test scores, but in principle, there is no reason why the value-added statistical methods cannot be applied to other student outcomes. Including attendance and credit accumulation has the potential to produce a broader and richer portrait of school quality for middle and high schools, where we applied this method. Our school VAM measures are now based on multiple tests as well as these non-test measures of student performance.

The report's findings on school-level value-added in high schools have particular relevance to the development of a formula for calculating team-based performance bonuses for Promise Readiness Corps (PRC) teachers who successfully deliver cohorts of entering 9th graders to the 11th grade. The model used to analyze the performance of PRC teachers will resemble quite closely the

VAMs used to evaluate school performance. Each PRC team will receive an estimate of its ability to improve student test scores in the same subjects analyzed in this report. These estimates will be weighted and combined to form a measure of PRC effectiveness which will enter the calculation of their bonus formula.

This report also analyzes four methodological issues: (1) we measure the effect of accounting for students' race/ethnicity on the value-added measures of schools and teachers; (2) we compare results of school VAMs for middle and high schools when using student baseline scores from the prior year versus baseline scores from the year prior to entering the school, with the aim of shedding light on the validity of VAMs for elementary grades where "pre-entry" scores are unavailable; (3) we discuss a method to account for differential dropout rates in high schools and describe diagnostic analyses suggesting the extent of the differential attrition problem; and (4) we further examine the consequences of accounting for three or more baseline achievement scores in a VAM rather than two baseline scores.

Subsequent reports will update and expand the existing findings. By March 1, 2011, we will develop composite measures that aggregate information from multiple subjects and assessments. As observational data become available, we plan to examine relationships among VAM results, student perceptions data, teacher working conditions survey results, and RISE classroom performance evaluations. We will also explore the possibility of using results from schoolwide working conditions surveys as additional controls in teacher-level value-added estimates.

This report is structured as follows: Chapter II describes the data used, including the student, teacher, and school samples, and the outcome measures; Chapter III describes the teacher VAMs and discusses the overall results and sensitivity analyses; Chapter IV describes school VAMs, results, and sensitivity analyses; and Chapter V discusses future work.

II. DATA AND SAMPLE

The value-added approach utilizes multiple years of achievement data on individual students linked to teachers, courses, and schools. PPS provided Mathematica with data from its Real-Time Information (RTI) database containing these links. This section describes the data, focusing on the size and characteristics of the samples as well as the quality of data linkages.

A. Students

The RTI database contains information on test scores, race, attendance, credits, whether a student is gifted, disability status, English Language Learner classification, age, and eligibility for a free or reduced price lunch. Students were included in the sample if we had data on their prior year PSSA scores and we could link them to teachers through the RTI database. We did not see evidence suggesting that student identifiers were not unique or that assessment scores were missing when they should be present. Nevertheless, 22% percent of the eligible sample must be excluded due to nonexistent 2009-10 or 2008-09 year PSSA scores. There were 11,663 students in the analysis sample, compared to 14,914 students in grades 4 to 11 in Pittsburgh schools in 2009-10. Table II.1 describes characteristics of students included in the analyses and students who could not be included, in grades four through eight.

Table II. 1. Descriptive statistics on the analysis sample in grades 4- 8, 2009- 10

	Students with PSSA Math and Reading Scores in 2009-10 (n=9139)	
	Included Sample	Excluded Sample
PSSA Math (z-score in 2009-10)	**-0.40**	**-0.72**
PSSA Reading (z-score in 2009-10)	**-0.32**	**-0.61**
African-American	**0.56**	**0.63**
White	**0.36**	**0.25**
Asian	**0.02**	**0.04**
Hispanic	**0.01**	**0.03**
Other	0.05	0.05
Meals Program	**0.74**	**0.70**
Total Students	7987	1152

Note: Bold indicates statistically significant differences at the 5 percent level.

Across schools in elementary and middle grades, the lowest percentage of included students was 73 percent in grades 4-5 and 71 percent in grades 6-8. Excluded students in high-school grades likewise were, on average, more disadvantaged than included students.

B. Teachers

Our estimates for the 2009-10 year apply to 646 teachers, an increase of more than 40 percent from our August report (Lipscomb, Gill, and Booker, 2010). The teacher value-added models cover PPS teachers assigned to courses that can be connected to assessment outcomes, including PSSA, CBA, PSAT, and SRI. The eligible sample included 1,976 total staff teaching students in 2009-10. We generate teacher VAM estimates in each subject for those successfully matched with more than

10 students in the analysis sample.[2] 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; Potamites et al., 2009a). Each teacher has at least one value-added estimate based on students taught in 2009-10. Of the 646 teachers, 494 teach in grades 4 to 8 and 152 teach in high school. The previous report included 452 teachers of math and reading in grades 4 to 8. Most of the increase in sample size results from extending teacher VAMs to grades 9 to 11 and to including middle school assessment outcomes for science and U.S. history. The addition of more assessments has also made it possible to produce more than one value-added estimate for many teachers, because we have now incorporated more than one student assessment into VAMs in some grades and subjects.

C. Schools

Analyses using 2009-10 data covered 22 elementary (K-5), 18 K-8, 12 middle (6-8), and 11 high schools. K-8 schools are given separate value-added estimates for grades 4-5 and grades 6-8. Thus they are compared to both elementary and middle schools. As with the student identifiers, the quality of student-school linkages appears to be high in the RTI database, permitting us to develop measures that account for students who attend multiple schools in a given school year.

D. Outcome Measures

As Table II.2 indicates, the VAMs included in this report now incorporate the great majority of student assessments that are under consideration for possible use in VAMs. An "x" in the table indicates that the assessment exists in that grade level and is being used; an "o" indicates that it exists but is not yet in use in the VAMs for this report; an empty cell indicates that the assessment is not given at that grade level. In upcoming work we will seek to add the 11th-grade assessments to the ones examined here.

In general, we use end-of-year test scores as the outcome measure of interest in each VAM. If we have data on the scores of a mid-year administration of one of these tests in addition to an end of year score, we use as our outcome measure the average of these two scores with the end of year score receiving twice as much weight as the mid-year score. We use the same approach at the middle school level for outcomes, such as the SRI and CBA, which are measured multiple times during the school year. The rationale behind including information on multiple scores is to reduce the measurement error present if only one administration of the test is used. The end-of-year score receives higher weight because the teachers and schools have had a longer time to influence the outcomes of students which means relatively more information on teacher and school effectiveness is contained in the end-of-year scores.

All of the identified assessments in Table II.2 are used in both school-level and teacher-level VAMs in this report. School-level VAMs for grades 6-12 also include estimates for a school's value added in terms of student attendance and credit accumulation. Those outcomes are not included in teacher-level VAMs because we do not believe they can be reliably attributed to individual teachers.

[2] A successful match occurs when a student-subject observation matches with a teacher who is listed in the course schedule as being the one who taught the student that particular course.

Table II.2. Available test scores by subject and grade

Column1	3	4	5	6	7	8	9	10	11	12
PSSA Reading	x	x	x	x	x	x			x	
PSSA Writing			x			x			x	
PSSA Math	x	x	x	x	x	x			x	
PSSA Science		x				x			o	
CBA Math				x	x	x				
CBA Algebra I							x			
CBA Algebra AB-BC							x			
CBA Geometry								x		
CBA Algebra II									x	
CBA English				x	x	x	x	x		
CBA Earth Science				x						
CBA Life Science					x					
CBA Biology							x			
CBA Chemistry								x		
CBA Physics						x			o	
CBA Civics							x			
CBA History						x		x		
SRI				x	x	x	x	x	o	o
PSAT Reading								x		
PSAT Math								x		

Note: Cells marked with an "x" are test scores currently used in the analyses. Cells marked with an "o" represent scores available but not currently in use. 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.

For purposes of developing a formula for performance-based bonuses for teams of Promise-Readiness Corps teachers in Pittsburgh's eight comprehensive high schools, PPS and PFT would like to include value-added estimates that, where possible, are benchmarked against historical performance district-wide rather than re-normed each year. Benchmarking against historical scores enables comparisons to be made between the current performance of schools with past performance. This allows for the possibility that all schools in the district may be improving or declining in quality simultaneously. When test scores are re-normed to district-wide averages each year, schools are competing directly with each other rather than with their historical performance.

PSAT and SRI tests are nationally normed and designed to have consistent scales from one year to the next, which allows them to be used to capture district-wide performance changes over time. We can therefore compare these scores to historical district-wide averages when using them in the PRC bonus calculations. In Pittsburgh, PSAT scores in fact show a very high level of consistency from 2008-09 to 2009-10. (We cannot examine the issue for SRI scores because we have only one year of data for those.) The key questions about historical benchmarks thus relate to whether other 9th and 10th grade outcome measure such as CBAs, attendance, or credits can also reliably be compared to historical district-wide averages.

To determine whether an outcome measure is stable over time we cannot simply compare the average score of students in the district from one year to the next. This is because it is possible for the schools and/or students to be improving in achievement over time. If this were the case the average score would be increasing not because the test is unreliable but because students are realizing achievement gains. On the other hand, we would not want the re-norming of an unstable test score to make it appear that there were achievement gains over time when none existed. Therefore we need a known stable benchmark with which we can compare the changes in our outcome measure of interest. Because they are normed nationally, the PSAT scores in Math and Reading provide such a benchmark measure.

For an outcome to be stable, we require that changes in the distribution of its scores be similar to the changes in distribution of PSAT scores. This requirement relies on the assumption that changes in PSAT scores are strongly correlated with changes in achievement as measured by other outcomes over time. This is likely to be true for English and Math CBAs, but less so for Science CBAs, History CBAs, attendance, and credits. We compare the changes in the CBA, attendance, and credits outcomes of students between 2008-09 and 2009-10 with the changes in PSAT scores in Tables II.3 and II.4.

Table II.3 Consistency over time of 9th grade outcome measures

Outcome	Mean, 2009-10	Standard Deviation, 2009-10	Mean, 2008-09	Standard Deviation, 2008-09	Mean Difference in 2008-09 SD Units	p-value on Test of Equivalence of Means	p-value on Test of Equivalence of SDs
December/January							
Algebra I CBA	0.51	0.23	0.53	0.18	-0.13	0.01	0.00
Biology CBA	0.53	0.17	0.51	0.17	0.08	0.03	0.88
Civics CBA	0.59	0.21	0.55	0.20	0.20	0.00	0.03
English I CBA	0.64	0.16	0.65	0.15	-0.05	0.22	0.06
May/June							
Algebra I CBA	0.46	0.15	0.45	0.17	0.08	0.07	0.00
Biology CBA	0.46	0.18	0.43	0.18	0.17	0.00	0.70
Civics CBA	0.64	0.18	0.62	0.21	0.07	0.08	0.00
English I CBA	0.67	0.17	0.67	0.17	0.02	0.69	0.23
Total Credits	7.92	1.49	7.51	1.60	0.25	0.00	0.00
Attendance	0.93	0.11	0.92	0.13	0.11	0.00	0.00

Note: Samples for CBA assessments include students in grades 9 or 10. Attendance is measured as the fraction of school days for which a student is present

Table II.4 Consistency over time of 10th grade outcome measures

Outcome	Mean, 2009-10	Standard Deviation, 2009-10	Mean, 2008-09	Standard Deviation, 2009-10	Mean Difference in 2008-09 SD Units	p-value on Test of Equivalence of Means	p-value on Test of Equivalence of SDs
December/January							
Chemistry CBA	0.44	0.14					
English 2 CBA	0.70	0.20	0.69	0.20	0.02	0.65	0.32
Geometry CBA	0.54	0.18	0.55	0.16	-0.06	0.18	0.00
History CBA	0.58	0.20	0.57	0.19	0.07	0.16	0.09
May/June							
Chemistry CBA	0.49	0.18					
English 2 CBA	0.68	0.17	0.63	0.18	0.29	0.00	0.13
Geometry CBA	0.44	0.19	0.52	0.18	-0.40	0.00	0.22
History CBA	0.52	0.21	0.44	0.21	0.39	0.00	0.68
PSAT Math	38.66	10.46	39.15	10.68	-0.05	0.22	0.43
PSAT Reading	36.36	11.31	35.95	11.56	0.04	0.34	0.42
Total Credits	8.07	1.56	7.68	1.61	0.25	0.00	0.19
Attendance	0.93	0.12	0.92	0.12	0.11	0.00	0.72

Note: Samples for CBA and PSAT assessments include students in grades 9 or 10. Attendance is measured as the fraction of school days for which a student is present.

Neither the PSAT score means nor the standard deviations appear to change between 2008-09 and 2009-10. The p-values of the hypothesis tests that the PSAT means and standard deviations are equal are all above .05, which means we cannot reject this hypothesis at the 5% significance level. Because the PSAT score distribution appears to be stable between 2008-2010, we would expect other reliable measures to also be stable over this time period. The only other measures which satisfy these stability criteria are the 9th grade mid- and end-of-year English I CBA, the mid-year 10th grade History CBA, and the mid-year 10th grade English II CBA. The means and standard deviations of these outcomes are not significantly different from each other between 2008-09 and 2009-10. Nonetheless, it is possible that the consistency of these measures was a one-time statistical fluke. Unless PPS has information suggesting that those assessments are likely to have consistent scales in the future, we do not recommend using a historical baseline in estimating VAMs for those results.

District-wide changes in PSAT scores might be used as a way of assessing the district's overall improvement in value-added. These district-wide changes might then be used to determine the total size of the bonus pool (and/or the number of teams eligible to win bonuses) available to Promise Readiness Corps (PRC) teams. This would allow the district to avoid imposing a "zero-sum game" in the awarding of bonuses: If performance across the district improved on the PSAT, then bonuses would potentially be available to most or even all of the PRC teams.

In future years, the PRC bonus formula might become less dependent on PSAT scores, because the implementation of statewide Keystone exams permits the calculation of value-added estimates for PRC teams relative to the performance of the state as a whole. The performance of the rest of the state would then become the benchmark, rather than the district's own performance on the PSAT in prior years. This, of course, is a policy decision for PPS and PFT to consider.

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

III. TEACHER VALUE- ADDED STATISTICAL MODELS AND RESULTS

A. Estimation Equation and Control Variables

The basic value-added model used to estimate teacher effects adjusts the scores of individual students for factors like student and peer background characteristics that are thought to correlate with achievement. The estimation equation is:

$$A_{i,j,y} = \sum_{j=1}^{J} \beta_j A_{i,j,y-1} + X_{i,j,y}\gamma + D_{i,j,y}\delta + e_{i,j,y}$$

where $A_{i,j,y}$ is the achievement test score for student i in subject j (i.e. math or reading) in year y and $A_{i,j,y-1}$ is the prior score for student i in subject j. $X_{i,j,y}$ is a set of control variables (described below), $D_{i,j,y}$ is a set of teacher variables, and $e_{i,j,y}$ is the error term. The coefficients β, γ, and δ capture the estimated relationships between outcome scores and each respective variable, controlling for other factors. The standard errors adjust for clustering of observations by student.

The research literature on value-added modeling typically assumes that $A_{i,j,y-1}$ captures all previous inputs into student achievement.[3] However, all tests will contain measurement error; including the prior score without correcting for measurement error can therefore bias the teacher dosage measures. To address this issue, most of our baseline models include two prior test scores—one in the same subject as the outcome of interest, and one in another subject.[4]

VAMs for middle school teachers are the exception to the rule about including two prior achievement scores. In these models, we control for one prior test score in the same subject as the outcome of interest, and one contemporaneous score in a different subject. We follow Coltfelter, Ladd, and Vigdor (2010) in using contemporaneous subject scores as controls when teachers are departmentalized to capture current-year influences that prior scores may not measure. At the middle school level, substituting one of the prior year controls for a contemporaneous control (e.g. prior year reading in a math VAM) improved the model's explanatory power without loss of sample size. High school models could use the same strategy, but we determined that using contemporaneous scores resulted in a smaller proportion of students in the analysis sample than using two prior scores as controls (because many high school students were not taking the same courses at the same time). Section E of this chapter shows results from a sensitivity analysis that compares both types of models at the high school level.

[3] This assumption simplifies the estimation of the general value-added model from theory. McCaffrey et al. (2009) provide a brief description with references to detailed derivations in several other frequently cited studies. The model also allows test score gains to be modeled as a function of current year factors only. We include several prior year variables, thereby relaxing this assumption somewhat.

[4] An alternative way to address measurement error is to use a two-stage least squares approach, where the student's prior test score in the other subject serves as an instrumental variable for the prior same-subject test score (Potamites et al., 2009a).

Test scores are normalized using the annual state or district average and standard deviation in each subject and grade level. 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. For each subject and grade, we subtract the average score from individual scores and then divide by the standard deviation of scores before estimating the model. Expressing scores like this allows us to interpret above-average scores in terms of how close to the average most students tended to fall.

We use a dosage approach in constructing the teacher variables as a flexible way of accounting for students who learn from more than one teacher in a given subject in a year and for students who are enrolled for only part of the year. The vector $D_{i,j,y}$ includes one variable for each teacher in the respective models. Each variable equals the fraction of the year student *i* was taught by a teacher in subject *j*.[5] The dosage value of any element of $D_{i,j,y}$ is zero if student *i* was not taught by that teacher. Because teachers are unlikely to have an appreciable educational impact on students who are enrolled only briefly, the dosage variable is also set to zero for students spending less than two weeks with a teacher and to one for students spending all but two weeks with a teacher. Students spending more than two weeks and less than all but two weeks are attributed to each teacher for the proportion of the year spent with that teacher. We also include a "residual teacher dosage" term that equals $1\text{-}\Sigma_T D_T$, where $\Sigma_T D_T$ for a given student is the sum of dosage variables across all teachers. For some students, we have information on their teachers for only part of the year (e.g., if they enter PPS mid-year). The residual teacher dosage variable accounts for the remaining time. By including this term in the VAM and by excluding the constant, the teacher value-added estimates can be measured relative to the average teacher. These measures are elements of the vector δ, which are the coefficients on $D_{i,j,y}$.

The vector $X_{i,j,y}$ includes several control variables. These variables are chosen as factors outside of the teacher's control so as to isolate their effect on student achievement. We normalize each variable by subtracting the mean and dividing by the standard deviation in each year.[6]

- Gender and race/ethnicity (white, African-American, Asian, Hispanic, other)
- Free or reduced price meals status
- English language learner (ELL) status
- Age
- Indicator of whether student is behind grade for age
- Variables for an identified disability and for gifted without disability
- Variables for changing schools at least once since beginning elementary or middle school

[5] The dosage is split between two teachers in the event a student appears to take multiple courses in the same subject at the same school.

[6] This normalization is done based on the mean and standard deviation from the analysis sample. In the future, we may explore adjusting these variables for differences relative to the state average as well since PPS scores are expressed in terms of the state score distribution but the characteristics of Pittsburgh students, like their demographics, differ from the state with respect to some variables.

- Separate variables for the proportion of prior year excused absences, unexcused absences, and days suspended
- A variable indicating prior year district membership the entire school year
- Classroom average shares by gender, race-ethnicity categories, meal program status, ELL status, disabled, and gifted (except in grades 4 and 5)[7]
- Class size and classroom averages for each of the following: prior year PSSA math and reading test scores, prior year proportions of excused and unexcused absences, prior year proportion of days suspended, and the proportion of students who were enrolled in PPS the entire prior school year (except in grades 4 and 5)
- Variables for student grade levels and their interaction terms with a student's prior year test scores in both subjects (for models that include multiple grade levels)

Some of these variables are correlated. Including related variables in the model does not interfere with the model's capability to estimate teacher effects consistently. In fact, it typically improves estimates so long as both control 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. For example, we find that students in the meals program tend to have lower enrollment in programs for gifted students and are more frequently absent for unexcused reasons. Suspensions tend to be positively related to both excused and unexcused absences.

To recognize that disability conditions vary widely across students, all of our VAMs now include control variables that distinguish most student disabilities at the category of disability level. The category-specific variables we include are for specific learning disabilities, speech and language impairments, mental retardation, emotional disturbance, and autism. While each of these disabilities includes at least one percent of the total enrollment, the remaining eight categories—with the exception of other health impairments—occur at much lower frequency. In some cases, they do not even identify a single student at each grade level. Our approach is to combine these remaining eight categories into two control variables for the analyses. The first includes physical/sensory disabilities (i.e. orthopedic impairments, hearing impairments including deafness, visual impairments including blindness, and deaf-blindness). The final category incorporates traumatic brain injury, developmental delay, and multiple disabilities into the other health impairments category. Although the coefficient estimates for each disability category vary by subject and grade level, they tend to be largest for specific learning disabilities and mental retardation. Coefficients on emotional disturbance and the category that includes other health impairments tend to be statistically significant as well, though somewhat smaller in magnitude.

The classroom variables listed above help control for peer influences in middle and high school grades. When classrooms are self-contained, as in elementary grades, it can be difficult to separately identify these influences from the contribution of the teacher. As a result, we must exclude these variables from models for teachers and schools in grades 4 and 5. Separately identifying teachers and peer influences is much less of an issue in secondary grades because teachers have several

[7] The classroom variables for a student use the classroom associated with the highest teacher dosage value.

classrooms of students. We are exploring the application of classroom-level variables to elementary teachers in the 3-year models, which permit examination of differences in classroom composition even for elementary teachers with self-contained classrooms.

We assume that the assignment of students to teachers and schools is functionally random conditional on the factors described above. To the extent that this assumption fails, VAM estimates may suffer from bias due to non-random assignment (Rothstein 2010). Fortunately, the degree of bias may not be large. Kane and Staiger (2008) were able to compare teacher VAM estimates in Los Angeles under a typical situation where principals assigned students to teachers to VAM estimates in the following year where principals assigned students to classrooms but then randomly assigned teaching assignments. They found that variation in teacher VAM scores before random assignment was a positive and significant predictor of achievement differences when classrooms were assigned randomly.

A limitation of value-added models is that they can only incorporate factors affecting student achievement growth for which data are available. For example, we currently lack information on other interventions that students might be participating in (during the school year or in summer school) that may help raise their test scores. If some students are receiving additional instruction in reading or math, any effect of that additional instruction would be mistakenly attributed to the classroom teacher. Because PPS plans to keep centralized data on participation in interventions, future VAM estimates will be able to factor out the effects of the interventions along with the other characteristics included in the model.

We also lack information about variations in the number of hours of instruction in particular subjects. The models do not currently account for the fact that some teachers or schools may spend more time, for example, on social studies than do other teachers or schools. To the extent that time allocations are determined by teachers or schools within the context of a fixed instructional day, it may be appropriate to ignore these differences in the VAM calculations, because the allocation of time is part of what determines effectiveness.

Ignoring differences in available resources is potentially more problematic if the differences 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. Similarly, we do not adjust results based on differences in teacher experience or teacher salary. PPS may wish to consider future supplemental analyses that account for these kinds of differences in resources across schools and teachers, thereby producing complementary VAM estimates that implicitly seek to compare value-added per dollar or value-added per hour.

B. Shrinkage Procedure

After estimating the value-added model, we use a shrinkage procedure to minimize the possibility that teachers with imprecise estimates (e.g., those with relatively small numbers of students) are over-represented among high-performing and low-performing teachers. The point of using this approach is to avoid overestimating the importance of teacher effects that we might reasonably suspect to be noisy. Shrinkage procedures are used regularly in value-added applications in education.[8] The procedure weighs information from two sources: information available on a specific teacher and on all teachers. If a teacher estimate is based on a large number of students, more weight goes to the information about the individual teacher, and vice versa. The procedure assumes that a teacher is an "average" teacher until more precise evidence allows us to revise that assumption. Following the shrinkage procedure, we further address the potential for noise by dropping teacher effects based on 10 or fewer students, and then center the remaining estimates on a zero value. This final step means that the average teacher in Pittsburgh has a value-added estimate of zero.

C. Model Modifications for Outcomes Other Than Math and Reading

Unlike reading and math, social studies is not tested in multiple successive grades in Pittsburgh. Prior scores must therefore use different subjects. For 9th grade civics we use the 8th grade reading PSSA as the prior score control under the assumption that reading PSSA measures a prior skill that is as similar as possible to skills needed for social studies. Similarly, we use the 7th grade PSSA in reading for the CBA U.S. history outcome in grade 8. Both of these outcomes include a score control in a different subject as well.

Prior year scores in science for science outcomes are sometimes, but not always, available. We include a control for the prior year score when one is available (e.g. control for life science CBA in grade 8 PSSA science model) and use a similar strategy as outcomes in social sciences when one is not available. For example, in the case of the PSSA science in grade 4, we control for grade 3 PSSA scores in math and reading as we do for other VAMs in elementary grades. Finally, we treat the PSSA writing assessment like the PSSA reading in that it applies directly to reading teachers. As such, we use the use the prior year reading score as the same-subject control in models describing writing performance.

Attendance and credits earned are examined in the school (or PRC cohort) value-added models (described in Chapter IV), but not in the teacher value-added models, because we assume they cannot be attributed to individual teachers.

D. Results

The results for elementary school teachers for one- and three-year models are displayed in Table III.1. Overall the three-year model uses more data on students to estimate the teacher VAMs which results in lower average standard error estimates and a higher fraction of teachers statistically distinguishable from average. This general outcome is consistent with what we found in the August report (Lipscomb, Gill, and Booker, 2010). This comes at the cost of producing estimates for fewer teachers, however, since we require teachers to have three years of data under the extended model. As indicated in Table III.1, the number of teachers for whom three-year estimates can be calculated is typically slightly less than half the number of teachers for whom one-year estimates can be calculated.

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

On average, using a 95 percent confidence interval, we can distinguish 35% of teachers from average in the one-year model and 45% in the three-year model. The year-to-year correlations of teacher effects are reasonably high and positive, with most correlations significantly different from zero. The within-teacher correlations across outcome measures are displayed in Appendix Table A.1. Overall the correlations are positive, indicating that if an elementary school teacher is successful at improving student test scores in one subject, that he or she is likely to be successful in other subjects as well.

Table III.1. Teacher VAM results for grades 4 to 5, by outcome

Outcome	Grades	Adj. R-squared	Students	Teachers	SD of Teacher Effects	Mean Standard Error	Statistically Significant Effects (95% CI)	Year-to-Year Correlation
One-Year Model (2009-10)								
PSSA math	4-5	0.78	3436	101	0.20	0.09	0.35	**0.53**
PSSA reading	4-5	0.74	3437	119	0.18	0.11	0.24	0.17
PSSA science	4	0.71	1737	44	0.18	0.09	0.30	**0.62**
PSSA writing	5	0.48	1682	58	0.37	0.16	0.43	**0.45**
Three-Year Model (2007-08 to 2009-10)								
PSSA math	4-5	0.76	10391	50	0.20	0.06	0.48	n/a
PSSA reading	4-5	0.74	10382	56	0.14	0.07	0.29	n/a
PSSA science	4	0.69	5066	21	0.14	0.06	0.38	n/a
PSSA writing	5	0.49	5132	30	0.33	0.14	0.50	n/a

Bold indicates that year-to-year correlations are statistically different from zero using a 95 percent confidence interval.

The results for middle school teachers are displayed in Table III.2. The table includes more outcome measures because CBAs in multiple subjects were given to middle school children. We have CBA and SRI data only for 2009-10, however, so year-to-year correlations are produced only for the PSSA scores. Overall the year-to-year correlations are large and positive. The three-year model at the middle school level provides increases in precision over the one-year model, similar to elementary school results. On average we can distinguish 35% of teachers from average in the one-year model and 48% of teachers from average in the three-year model. The within-teacher correlations across outcome measures, for teachers who have more than one student assessment associated with their work, are displayed in Appendix Table A.3. At the middle school level teachers become compartmentalized and are likely to teach only one subject. Therefore correlations are only available across outcome measures in the same subject area.

Table III.2. Teacher VAM results for grades 6 to 8, by outcome

Outcome	Grades	Adj. R-squared	Students	Teachers	SD of Teacher Effects	Mean Standard Error	Statistically Significant Effects	Year-to-Year Correlation
One-Year Model (2009-10)								
PSSA math	6-8	0.83	4780	93	0.17	0.07	0.38	**0.64**
PSSA reading	6-8	0.77	4779	140	0.11	0.08	0.13	**0.43**
PSSA science	8	0.77	1514	29	0.11	0.08	0.10	0.24
PSSA writing	8	0.62	1597	44	0.25	0.13	0.34	**0.78**
CBA math	6-8	0.52	3668	82	0.40	0.14	0.51	n/a
CBA English	6-8	0.60	3655	99	0.22	0.13	0.22	n/a
CBA earth science	6	0.69	1365	33	0.47	0.13	0.64	n/a
CBA life science	7	0.71	1383	31	0.34	0.13	0.48	n/a
CBA physics	8	0.67	1438	28	0.31	0.13	0.32	n/a
CBA US history	8	0.69	1389	26	0.38	0.12	0.50	n/a
SRI	6-8	0.65	3310	117	0.24	0.13	0.32	n/a
Three-Year Model (2007-08 to 2009-10)								
PSSA math	6-8	0.82	15318	50	0.15	0.04	0.52	n/a
PSSA reading	6-8	0.77	15277	84	0.11	0.05	0.36	n/a
PSSA science	8	0.74	5007	16	0.13	0.05	0.56	n/a
PSSA writing	8	0.60	5082	26	0.29	0.08	0.73	n/a

The results of the teacher VAMs at the high school level are displayed in Table III.3. These results include only one year of teaching; most of the assessments have not been in place for three years. With one year of teaching, we can distinguish 16% of teachers from average across all assessments. The year-to-year correlations in teacher effects appears to be somewhat lower than at the middle or elementary school level, but care should be taken in interpreting these results. We have data on many fewer teachers at the high school level than we do at the lower grades, which leads to imprecise measures of the year-to-year correlations of effects. We do not report a table of within teacher correlations of effects across outcome measures due to the small sample sizes associated with these statistics. When we use additional data on teachers in the two year model the precision of our estimates increases and we are able to distinguish a larger fraction of teachers from average.

Table III.3. Teacher VAM results for grades 9- 11, by outcome and year

Outcome	Adj. R-squared	Number of Students	Number of Teachers	SD of Teacher Effects	Mean Standard Error	Fraction Statistically Significant Effects	Year-to-Year Correlation
Grade 9 (2009-10)							
CBA Algebra AB-BC	0.53	255	11	0.42	0.65	0.09	
CBA Algebra I	0.48	544	16	0.11	0.18	0.00	0.14
CBA Biology	0.58	1034	21	0.55	0.14	0.62	0.77
CBA Civics	0.54	954	21	0.23	0.16	0.24	0.65
CBA English I	0.54	888	24	0.27	0.19	0.17	-0.24
SRI (Dec/May)	0.68	916	32	0.37	0.18	0.25	
Grade 10 (2009-10)							
CBA Chemistry	0.53	775	17	0.42	0.15	0.35	
CBA English II	0.57	761	21	0.25	0.16	0.19	
CBA Geometry	0.67	657	22	0.39	0.18	0.41	
CBA History	0.45	665	18	0.33	0.17	0.33	
SRI (Sept)*	0.49	587	25	0.17	0.18	0.04	
PSAT Math*	0.71	1000	39	0.13	0.14	0.00	0.05
PSAT Reading*	0.63	998	33	0.02	0.07	0.00	0.37
PSAT Writing*	0.59	982	33	0.08	0.11	0.00	0.73
Grade 11 (2009-10)							
CBA Algebra 2	0.62	583	27	0.58	0.24	0.52	
PSSA Math	0.69	622	25	0.18	0.16	0.08	
PSSA Reading	0.64	725	31	0.05	0.12	0.00	
PSSA Writing	0.37	700	31	0.27	0.24	0.10	
Two Year Model (2008-09 and 2009-10)							
CBA Algebra I	0.55	1490	7	0.27	0.11	0.57	
CBA Biology	0.58	1801	10	0.42	0.10	0.80	
CBA Civics	0.51	1992	14	0.26	0.10	0.43	
CBA English I	0.52	1844	10	0.15	0.10	0.30	
PSAT Math*	0.70	2133	18	0.08	0.09	0.06	
PSAT Reading*	0.63	2116	14	0.07	0.08	0.07	
PSAT Writing*	0.60	2073	14	0.11	0.09	0.21	

E. Additional Modifications and Sensitivity Analyses

In this section, we consider variations on the benchmark VAMs described above both to evaluate their quality of fit to the data and to weigh the benefits and costs of alternative approaches. In the previous report, we conducted a sensitivity analysis that averaged impact estimates for each teacher across multiple student years (i.e., multiple years of teaching). We found that this improved the precision of the results, making it possible to identify high- and low-performing teachers with greater confidence. Using a three-year rolling average involves a tradeoff in the timeliness of the estimates, so PPS may want both one-year and three-year estimates. The one-year estimate would provide a snapshot of the most recent year of teaching while the three-year estimate would provide a more stable long-term assessment of teacher performance that is likely to be better suited to high-stakes decisions.

Below we describe additional sensitivity analyses related to the inclusion of additional baseline scores (beyond the first two); the omission of control variables for student race/ethnicity; and the inclusion of school effects.

Sensitivity to inclusion of additional scores for students as control variables

We attempted in August to measure gains in estimate validity and reliability versus costs in terms of reduced sample by adding additional pretest controls for each *student* to teacher VAMs in grades 4 to 8 (Lipscomb, Gill, and Booker, 2010). Those analyses found very high correlations among teacher effects and minimal changes to precision, and so we concluded that using two baseline scores did not perform measurably worse relative to a model with more control variables and it preserved students in the sample who do not have more than two prior scores. Nonetheless, there remains a possibility that a small amount of bias is introduced by having two prior scores rather than three or more, so we examined the issue further

We have now considered two additional possibilities related to the general issue of involving more than two scores while maintaining sample size. The first was the potential to include an indicator variable to designate when students are missing a prior score (and replacing the missing score with a specific value like -99). We concluded, however, that this model may actually introduce substantial bias (see Jones 1996).

The second strategy was to include a contemporaneous score in a different subject along with the two prior scores, for teacher VAMs in departmentalized grades. We expected that this approach might improve the validity and reliability of the estimates with minimal loss of sample size. In middle school grades, this appears to be the case. However, we found at the high school level that problems associated with missing data and students taking courses out of grade often led to dramatically reduced sample sizes when contemporaneous scores are added.

Table III.4 illustrates this problem by presenting contrasting results for the 2009 CBA Algebra II and CBA English II. For the Algebra II VAM we used the PSSA reading as a contemporaneous control. Missing data on the PSSA exam caused us to lose approximately 100 students from our sample, reducing the size from 27 teachers and 583 in the main results to 24 teachers and 479 students. This loss in sample size also caused a reduction in the precision of our estimates; the average standard error increased from 0.24 in the main results to 0.27 in the baseline VAM in the current baseline specification. Ultimately, adding the additional contemporaneous score did not appear to affect the results much. The teacher effects were almost perfectly correlated with and without the additional score, and including the additional score did not show that we had falsely distinguished any teachers from average in the baseline specification.

When we performed the check for the English II exam using the CBA Geometry score as an additional control, the loss in sample size was even greater than in the previous specification. Due to missing data and a large number of students taking Geometry in the 9th or 11th grades, we did not have contemporaneous score data for many students from the original 10th grade English II CBA analysis. Only 555 students and 18 teachers were included in the English II CBA sensitivity analysis whereas 761 students and 21 teachers were included in the benchmark results. The average standard error of the teacher effects also grew relative to the benchmark model. This loss of precision contributed to identifying only 11% of teachers as distinguishable from average whereas we distinguished 19% in the main specification. As a result of the loss of sample size and precision from including a contemporaneous score for most of the outcomes of interest, and because the results are otherwise almost identical with and without the additional control, we chose to omit contemporaneous scores in the remainder of the analyses.

Table III.4. VAMs including an additional contemporaneous control score

	CBA Algebra II 2009		CBA English II 2009	
	Baseline	Additional Score	Baseline	Additional Score
1-Year Model (2009-10)				
Standard deviation of teacher effects	0.70	0.69	0.23	0.24
Mean standard error	0.27	0.26	0.19	0.19
Fraction distinguishable from average	0.58	0.58	0.11	0.11
Statistically above-average effects	0.25	0.25	0.06	0.06
Statistically below-average effects	0.33	0.33	0.06	0.06
Number of students	479		555	
Number of teachers	24		18	
Correlations Between Estimates				
Within teacher correlation	1.00		0.99	
False positive rate for above-average effects	0.00		0.00	
False positive rate for below-average effects	0.00		0.00	

Note: The score on PSSA reading is added as an additional contemporaneous control score in the Algebra II VAM. The geometry CBA is added as an additional contemporaneous control score in the English II CBA VAM. The samples consist of students in the Baseline and Additional Score regressions (the sample size is lower than in the main results because of missing data on the additional score in each VAM). The false positive rates assess whether the application of a 95 percent confidence interval to the baseline model identifies teachers who would not be identified using a 90 percent confidence interval and an additional contemporaneous score.

Exclusion of race/ethnicity coefficient

Our prior report (Lipscomb, Gill, and Booker, 2010) found that race-ethnicity was among the factors explaining a statistically significant part of the variation in student achievement in Pittsburgh. For example, the models predicted higher achievement scores for white students than for black students even controlling for other factors. Our baseline VAMs included race-ethnicity indicators because these variables improve the explanatory power of the models. We acknowledge, however, that the data only indicates that an achievement gap by race exists, not its source. If the estimates on these variables are measuring the influence of unobserved factors outside of classrooms and schools like differences in community background characteristics, including student controls for race-ethnicity provides a more rigorous design for estimating the contributions of teachers and schools. If race is a proxy for other kinds of disadvantage that we cannot measure, then excluding race from the VAMs would implicitly penalize teachers and schools with above-average proportions of African-American students (and conversely inflate the VAM estimates of teachers and schools with below-average proportions of African-American students).

If, on the other hand, the race-ethnicity controls are measuring the influence of factors under the control of the teacher or school—i.e., if African-American students are doing worse because their teachers and schools are not teaching them as well as they are teaching white students—then including race as a control variable would be a mistake.

We performed sensitivity analyses on the inclusion and exclusion of controls for race in the teacher VAMs to better understand the extent to which these variables affect the teacher estimates themselves. Table III.5 displays results for grades 4-8 (one-year model). The distribution and mean standard error of the estimates are indeed quite similar with and without the inclusion of race controls. In addition, similar fractions of teachers are distinguishable from average in the two specifications and the within-teacher correlations are at least 0.97 across subjects.[9]

Table III.5. Grades 4- 8 VAMs for grades 4- 8 that include and exclude race controls

	PSSA Math		PSSA Reading	
	Baseline	Excluding Race	Baseline	Excluding Race
Standard Deviation of Teacher Effects	0.19	0.18	0.16	0.15
Mean Standard Error	0.08	0.08	0.10	0.10
Statistically Significant and Above-Average	0.16	0.14	0.12	0.11
Statistically Significant and Below-Average	0.16	0.18	0.11	0.12
Within Teacher Correlations	0.98		0.98	

Note: Models include prior scores, student characteristics, and classroom characteristics. Models that exclude race controls omit them at the student and classroom levels. Statistically distinguishable is defined using a 95 percent confidence interval. The models include 192 teachers in math and 259 teachers in reading.

Excluding the race controls seems to have a comparably stronger effect on results at the high school level. Table III.6 shows findings for the Civics and English I CBAs. The within-teacher correlation of the estimated effects is as low as 0.90 and the rate at which teachers are falsely distinguished from average is as high as 50%.10 These statistics provide some evidence that—at least at the high school level—race variables should be included in the estimation to avoid biasing the teacher effect estimates.[11]

[9] Results for school VAMs were very similar, as were multi-year models for both teachers and for schools.

[10] Because the number of teachers distinguishable from average under the baseline model is relatively small, a small change in the number of teachers distinguishable from average when race controls are excluded can lead to a very large percentage of false positive results.

[11] The differing results by grade level do not appear to reflect small samples in some race-ethnicity groups, such as Asian or Hispanic students, that could be driving VAM estimates for teachers teaching those students.

Table III.6. High school teacher VAMs including and excluding race controls

	CBA Civics 2008		CBA English I 2009	
	Baseline	Excluding Race	Baseline	Excluding Race
1-Year Model (2009-10)				
Standard deviation of teacher effects	0.22	0.21	0.27	0.26
Mean standard error	0.15	0.13	0.19	0.17
Fraction distinguishable from average	0.17	0.30	0.17	0.29
Statistically above-average effects	0.13	0.17	0.08	0.17
Statistically below-average effects	0.04	0.13	0.08	0.13
Number of students	1045		888	
Number of teachers	23		24	
Correlations Between Estimates				
Within teacher correlation	0.90		0.95	
False positive rate for above-average effects	0.25		0.25	
False positive rate for below-average effects	0.50		0.00	

Note: The comparison and baseline groups include the same sets of students. The models excluding race controls exclude the controls both at the student and classroom level. The false positive rates assess whether the application of a 95 percent confidence interval to the model excluding race identifies teachers who would not be identified using a 90 percent confidence under the baseline model.

School variables in teacher VAMs

One way to account for unobserved characteristics of schools that can potentially affect student achievement growth is to control for the school the student attends by including school indicator variables, also known as school fixed effects. Using this approach means that all comparisons are made within schools, with differences in student achievement growth across schools absorbed by the school fixed effects. Alternatively, excluding school-level variables from the regressions means that the effect that schools have on student growth is included in the teacher effects. This means that teachers teaching at good schools will have an advantage in the sense that their estimated effects will be higher than similar teachers who teach at lower quality schools.

Including school fixed effects may not necessarily be desirable, however, when the results of the teacher VAMs are being used to evaluate teachers and determine performance bonuses. When we include school effects we are implicitly comparing teachers directly to other teachers within the same school rather than to other teachers in the district. Therefore an above average teacher who is not quite as good as his/her colleagues at a very high performing school could actually be estimated to be below average in a model with school fixed effects. This would be especially undesirable if part of the unobservable reason a school is high performing is the positive effect generated by having many good teachers.

An alternative method to including school fixed effects that still accounts for the impact schools have on student achievement is to include school level observable characteristics in the VAM. Examples of such school level observable characteristics are the fraction of students eligible for free or reduced price lunches or the fraction of students with disabilities. Including school observables separates some of the school influences on student achievement from the teacher VAM estimates to give a more accurate depiction of teacher effectiveness. However, other school level factors could still be influencing student achievement and therefore biasing teacher effects. The choice about which school level variables (if any) will be determined in part by the previously mentioned tradeoffs but also by the availability of sufficient data to identify the coefficients on the school variables.

With only one year of data we cannot separately identify school level variables from teacher level variables. Because our goal is to estimate a teacher VAM, in this case where we have only one year of data we need to exclude all school level variables form the regression.

If we include more than one year of data in the VAM it becomes possible to identify both school fixed effects and school observable characteristics. Depending on the amount of variation in the data, however, identification may be weak. School fixed effects would be identified by teachers transferring from one school to another to another across years. Given the short nature of our panel on teachers in Pittsburgh, it is unlikely that we would have enough teacher transfers for reliable identification. The effect of school observables, by contrast, would be identified by changes in the characteristics within a school over time. However, a variable like the fraction of students with disabilities is unlikely to vary much from year to year, leading to problems with weak identification.

It should be noted that once we control for student characteristics, prior scores, classroom average characteristics, and teacher effects, it is possible that schools do not have a substantial additional impact on student achievement. We can get an idea of the magnitude of the impact of omitting school effects in our teacher VAMs by examining the change in standard deviation of teacher effects after subtracting off the within-school average teacher effect from our estimates. We therefore took the VAM estimates from the basic one-year model estimated earlier and subtracted the within-school average teacher effect from each teacher effect.[12] Mechanically, this has the intended effect of comparing teachers to the average teacher in their school. Table III.7 (reproduced from our August report) compares the resulting distributions "with" and without school effects for 2006-07, 2007-08, and 2008-09.[13]

Table III.7. Standard Deviation of Teacher VAM Estimates "with" and without School Effects

	Math		Reading	
	No School Effects	"With" School Effects	No School Effects	"With" School Effects
2008-09	0.18	0.15	0.18	0.16
2007-08	0.19	0.15	0.14	0.12
2006-07	0.19	0.15	0.17	0.14

Note: Models without school effects compare teachers to the average teacher in Pittsburgh. Models "with" school effects compare teachers to the average teacher in their school.

The findings indicate that 65 to 80 percent of the variation in teacher effectiveness in Pittsburgh is within schools rather than across schools. In other words, the variation in value-added is larger

[12] We average effects for teachers who teach at multiple schools.

[13] "With" is placed in quotation marks because school effects are not actually included in the VAM, although subtracting the within school average teacher effect shares similar properties to the inclusion of school effects.

among teachers than it is among schools.[14] We see this because the standard deviation of the VAM estimates after adjusting for school effects is not much lower than it is without adjusting for school effects. Given the discussion above, the 20 to 35 percent of the variation that is across schools is best viewed as an upper bound on the amount of the variation attributed to teachers in our benchmark VAM that actually is due to unobserved school-level factors.

Table III.7 is consistent with prior research that finds greater variance in value-added at the teacher level than at the school level (Hanushek and Rivkin, 2010). For example, Aaronson, Barrow, and Sander (2007) find that the standard deviation of math teacher effects drops from 0.24 to 0.19 when school effects are included. Similarly, Kane and Staiger (2008) find that the standard deviation of math teacher effects drops from 0.231 to 0.219 in math and from 0.184 to 0.175 in reading.

Given the short length of time for which we observe teachers, we will not be able to include school fixed effects in the analysis. Before any results are released next spring, we will examine the extent to which sufficient variation exists in the school observables for us to include them in the model. We will also explore an alternative method of taking into account school observable characteristics. This method consists of first estimating the VAM without school observables, then regressing the teacher effects on the school observable characteristics, then using the residuals from this regression as our estimates of the teacher effects. We will explore the sensitivity of the results to the use of this method and to the inclusion of school observables directly by comparing these teacher effects to those estimated without school observables.

Ceiling effects and nonlinearities

Some teachers in Pittsburgh may have classes with unusually large or small fractions of high achieving students. The effects we estimate for these teachers in our VAMs will be unbiased as long as we adequately control for the prior year achievement scores of the students in the class. In our main VAM specifications we assume that a student's current year performance is a linear function of his or her prior year performance.

If a large fraction of students achieve a perfect score on the test, this can pose a problem for estimation of teacher effects because a student cannot improve upon a perfect score, so a teacher cannot be expected to increase this student's achievement. This violation of linearity will only pose a problem for the estimation of teacher effects if high achieving students are concentrated in certain classes.

Fortunately for our purposes, none of the assessments we examine appear to exhibit strong ceiling effects (with the exception of the attendance measure which will be discussed in the following chapter). Even if ceiling effects were to exist, some existing research (Koedel and Betts 2010) shows that moderate levels of these effects do not have a large impact on the teacher VAM estimates. Koedel and Betts used actual data from a school district and simulated different levels of test ceiling levels to examine the impacts on estimated teacher effects. They found that even when 15% of student test scores are at the maximum level, the correlation between teacher effects estimated in the restricted model and model with no ceiling effects is 0.97.

[14] Approximate standard errors for the VAM estimates with school effects are similar to those of models that exclude school effects.

Even if ceiling effects are not a problem in the data, it could still be possible that there is a nonlinear relationship between prior year and current year test scores. It could be the case that it is harder for a student near the high end of the distribution to improve his or her achievement relative to a student in middle of the distribution. If this were true it would introduce bias in our teacher effect estimates because we would over-predict the current year scores of the high achieving students (and under-predict the scores of the lower achieving students) as a result of the linearity assumption. The linearity assumption can easily be relaxed, however, by allowing current year test scores to be a polynomial function of prior year scores. We examined the changes in the estimates when we allow for current year scores to be a cubic function of prior year scores (results not reported). The teacher effects are very similar whether or not the cubic controls are included, with correlations often higher than 0.99. We did, however, find that in some cases the polynomial terms were significantly different from 0, indicating that there is in fact a non-linear relationship between current year and prior year scores. We will therefore include polynomial terms in all of our future VAMs (even though the bias introduced by excluding them is quite small).

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

IV. SCHOOL VALUE- ADDED STATISTICAL MODELS AND RESULTS

A. Estimation Equation for Calculating School Effects within Pittsburgh

The VAMs for estimating school effects resemble the teacher model described in Chapter II. The model used for school-level value-added in grades 6-10 is the following:

$$A_{i,j,y} = \sum_{j=1}^{J} \beta_j A_{i,j,yp} + X_{i,j,y}\gamma + D_{i,y}\delta + e_{i,j,y}$$

$A_{i,j,t}$ is the achievement test score for student i in subject j in year y and $A_{i,j,yp}$ is the score prior to entering the grade range, i.e., a fifth-grade score for students in grades 6-8 or an eighth-grade score for students in grades 9-10. We normalize test scores using the annual district-wide or statewide average and standard deviation in each subject and grade level. $X_{i,j,y}$ is the same set of control variables as used in the teacher model described in Chapter III, $D_{i,y}$ is a set of school variables, and $e_{i,j,y}$ is the error term.[15] The coefficients β, γ, and δ capture the estimated relationships between PSSA subject scores and each respective variable, controlling for other factors. Standard errors adjust for clustering at the student level.

The school dosage variables are calculated similarly to the teacher dosage variables. Each variable equals the fraction of the year student i was taught at a given school. The dosage value of any element of $D_{i,y}$ is zero if student i was not taught at that school. Because schools are unlikely to have an appreciable educational impact on students who are enrolled only briefly, the dosage variable is also set to zero for students spending less than two weeks enrolled at a school and to one for students spending all but two weeks enrolled at a school.

The dosage variables sum to one for students enrolled in PPS schools for the entire year. Like the teacher VAM, we include a "residual school dosage" term that equals $1-\Sigma_S D_S$, where $\Sigma_S D_S$ for a given student is the sum of dosage variables across all schools. The residual school dosage variable accounts for any part of the school year that students spend enrolled outside of PPS. By including this term in the VAM and by excluding the constant, the school value-added estimates can be measured relative to the average school in Pittsburgh. These measures are elements of the vector δ, which are the coefficients on $D_{i,y}$.

The school VAM makes the same assumptions as the teacher model and has similar limitations. Specifically, we assume that the assignment of students to schools is functionally random conditional on the included control variables. In addition, we interpret the estimates on the dosage variables as the actual contributions of schools to student achievement growth relative to the contribution of the average school in the district. Like the teacher VAM, we can only incorporate factors into the model that are observable in the data. School VAM estimates may suffer from bias due to omitted variables

[15] The school VAM controls for peer influences at the classroom level, as in the teacher model. We found that the VAM estimates were less stable when controlling for peer influences at the school level because of collinearity between the classroom peer effects and the school peer effects.

if they do not adequately control for the influence of any unobserved factors that vary across the communities that schools serve.

We apply the shrinkage procedure from Chapter III to ensure that schools with imprecise VAM estimates are not over-represented among high-performing and low-performing schools. The procedure works the same way as it does for teachers, by weighing information on a specific school and on all schools. If a school estimate is based on a large number of students, more weight goes to the information about the individual school, and vice versa. We then center the VAM estimates on a zero value.

The VAM for grades 4 and 5 differs slightly from the VAM described above, and more directly mimics the teacher VAM. Unlike in middle schools and high schools, in K-5 and K-8 schools there is no "pre-treatment" measure of student achievement—we have no information on students' achievement levels prior to entering kindergarten. As an alternative, the models for school value added in grades 4 and 5 rely on the preceding year's scores (from grades 3 and 4) for baseline purposes, just as the teacher models do. Later in the chapter we discuss sensitivity analyses at the middle and high school levels that examine whether using last year's score produces results similar to using pre-entry scores. We find that results using last year's scores are strongly correlated with results using pre-entry scores, suggesting that using last year's scores is a reasonable alternative method for grades 4 and 5. Nonetheless, this does not address the absence of any value-added measures for grades K-3. In the next section we discuss planned exploratory analyses that will attempt to assess early elementary achievement trajectories using DIBELS data from kindergarten for baseline purposes.

B. Possible Modification for Elementary Grades

With standardized testing beginning in grade 3, we have no information about value added during the first four years of schooling (grades K-3). This will improve slightly with the deployment of a standardized test at grade 2, but will still leave much of elementary-school instruction outside the realm of value-added modeling. We are therefore exploring another method of examining achievement trajectories in grades K-3, described here.

The Dynamic Indicators of Basic Early Literacy Skills (DIBELS) are a set of tests designed to be given to students between kindergarten and 6th grade to assess their progress in learning to read. We have data on DIBELS scores from 2004-2009 in Pittsburgh given to students between kindergarten and 3rd grade. We are planning to use this data to run school level VAMs to assess the effectiveness of schools at improving student reading ability through the 3rd grade. This will be an interesting addition not only to the PPS project but to the literature on VAMs in general, because the earliest most schools have standardized testing available is the end of 3rd grade.

We do not anticipate incorporating the DIBELS-based models into overall assessments of value-added in future years, because DIBELS results are too easily biased when high stakes are attached. They are administered and scored by individual teachers, creating a conflict of interest if stakes are attached. We therefore do not recommend using DIBELS results as baseline achievement measures for future analyses, whether for teacher- or school value-added estimates. (This might merit reconsideration if in the future the district found a way to have DIBELS assessments administered by central office staff rather than classroom teachers.) Nonetheless, an examination of DIBELS-based value-added retrospectively may provide useful information about ways to use PSSA scores in grades 3-5 to assess value added across all elementary grades.

In particular, using DIBELS data will allow us to compare the effectiveness of schools at improving student reading ability between kindergarten and 3rd grade with their ability to improve student outcomes in grades 4-5, and with their results in grade 3 as adjusted by observable student characteristics. This analysis could, in principle, allow us to refine the analysis of results in grades 3-5 so that it better captures K-5 performance.

More specifically, our exploratory, retrospective analysis will make use of several DIBELS assessments. The battery of DIBELS exams consists of five main tests given up to three times per year at various grade levels. The tests are: Initial Sound Fluency (ISF), Letter Naming Fluency (LNF), Phoneme Segmentation Fluency (PSF), Nonsense Word Fluency (NWF), and Oral Reading Fluency (ORF). We will use the scores on the beginning of kindergarten administrations of ISF and LNF as initial conditions and assume that they capture the ability of students prior to entering school. We will use the end of year ORF test in grade 3 as the outcome to assess school effectiveness at increasing student reading achievement in grades K-3. We will then correlate these estimated school effects with those calculated using PSSA scores to measure school effects in grades 4 and 5, and to cross-sectional estimates of 3rd-grade value-added. We will also use the end of year DIBELS assessments in kindergarten, 1st, and 2nd grade to estimate yearly VAMs for schools and examine how well these estimates are correlated across time.

Similar to our other analyses, we will standardize all scores to z-scores by subtracting the mean and dividing by the standard deviation. Since the outcomes of DIBELS tests are often of measured as the number of words a student gets correct in a given time period it is possible that students who are very proficient at reading can produce large outliers which will skew the score distribution. To prevent outliers from having large effects on the results, we will top-code student scores at 3 standard deviations.

C. Modification for Estimating Value- Added for Student Attendance

The analysis of attendance as an outcome variable differs from the other analyses because there is a substantial fraction of students that are at the upper limit of 100% (perfect attendance). Approximately 10% of students have perfect attendance in any given grade or year. If we were to estimate the VAM using OLS the resulting coefficient estimates would be biased by the fact that we are assuming a linear relationship between attendance and the covariates in the regression but the relation can no longer be linear once a student a student achieves perfect attendance.

To eliminate the bias associated with students having perfect attendance we use a Tobit model (Tobin 1958) to estimate the VAM. Let a_i^* be a latent variable measuring the ability of a student. We observe the attendance record a_i which is the fraction of school days that the student is present for. Let us assume that:

$$a_i = \begin{cases} a_i^* \; if \; a_i^* < 1 \\ 1 \; if \; a_i^* \geq 1 \end{cases}$$

Let us also assume that the relationship between a_i^* and the vector of covariates x_i is of the form:

$$a_i^* = \beta x_i + e_i$$

where e_i is a normally distributed error term with mean 0 and variance σ^2. The vector of coefficients β is estimated through maximum likelihood. The likelihood function in the Tobit model consists of two parts. The first part represents the likelihood that a value of a_i is observed that is less than 1, the second part represents the probability that the upper limit of attendance is reached. The likelihood function is of the form:

$$\prod_{i=1}^{N}\left[\frac{1}{\sigma}\phi\left(\frac{a_i-\beta x_i}{\sigma}\right)\right]^{1-I(a_i)}\left[1-\Phi\left(\frac{1-\beta x_i}{\sigma}\right)\right]^{I(a_i)}$$

where ϕ and Φ are the pdf and cdf of the standard normal distribution, $I(a_i)$ is an indicator variable that equals one if the student has perfect attendance, and N is the number of observations.

In the attendance VAMs we control for a student's prior attendance record along with the same additional variables included in the test-score VAMs (including two previous test scores). To allow for the effect of prior year attendance to be nonlinear because of the upper limit of perfect attendance, we add an indicator variable which equals one of the student had perfect attendance in the prior year to the set of covariates.

The VAM results presented in this report do not distinguish between excused and unexcused absences. Although excused and unexcused absences are given separate codes in Pittsburgh's data, initial examination of those data suggested that standards for determining whether an absence is excused or unexcused may vary over time and across the district. Overall rates of absence appear to be more stable over time than rates of excused or unexcused absences. We therefore use the overall rates in the current VAM estimates. Nonetheless, if PPS develops confidence in the consistency of standards for unexcused absences, future analyses could include only those, excluding the excused absences.

D. Results by Student Outcome

Results for Grades 4-5

The results of the school level VAMs on data from grades 4 and 5 are presented in Table IV.1. Similar to the results from the teacher VAMs, the school VAMs show increased precision when a multi-year model is used rather than a one-year model. On average 35% of school can be distinguished from average in the one-year model and 47% can be distinguished from average in the two-year model. The year-to-year correlations in school effects are generally around 0.5 and significantly larger than zero. Attendance data was only available to us for the 2009-10 school year, so the attendance outcome does not appear in the two-year models.

Table IV.1. School VAM results for grades 4 to 5, by outcome

Outcome	Grades	Adj. R-squared	Students	Schools	SD of School Effects	Mean Standard Error	Statistically Significant Effects	Year-to-Year Correlation
One-Year Model (2009-10)								
PSSA math	4-5	0.76	3461	39	0.14	0.06	0.26	**0.44**
PSSA reading	4-5	0.72	3462	39	0.09	0.06	0.18	0.11
PSSA science	4	0.71	1757	39	0.14	0.10	0.15	**0.36**
PSSA writing	5	0.47	1687	39	0.34	0.14	0.38	**0.57**
Attendance rate	4-5	n/a	3480	39	0.18	0.10	0.36	n/a
Two-Year Model (2008-09 to 2009-10)								
PSSA math	4-5	0.74	6970	39	0.14	0.04	0.41	n/a
PSSA reading	4-5	0.71	6966	39	0.08	0.04	0.23	n/a
PSSA science	4	0.69	3487	39	0.18	0.07	0.49	n/a
PSSA writing	5	0.49	3427	39	0.27	0.09	0.46	n/a

Note: An r-squared value is not calculated in the attendance rate VAM because it uses a different model specification, known as a Tobit model, to adjust for students with perfect attendance (that is, students who are top-coded). Approximately 10 percent of students in the sample have perfect attendance. Correlations in bold are statistically different from zero using a 95 percent confidence interval.

Results for Grades 6-8

The results from the middle school VAMs are displayed in Table IV.2. The number of schools varies across outcomes because some schools were missing data on CBAs. We again see increases in precision when an additional year of school data is used to estimate the two-year model. The year-to-year correlations in school effects are for the most part positive and significantly different from zero. They tend to be slightly higher than the correlations at the elementary school level. For the most part the fraction of middle schools distinguishable from average is similar to the fraction of distinguishable elementary schools.

The VAM for credits earned is unusual, in that 90% of schools can be distinguished from average based the credits VAM. This number should be interpreted with caution, however. It is possible that schools simply have different methods of assigning credits to classes and that some schools end up assigning more credits to students than others. The large standard deviation of school effects based on the credits outcome is consistent with this hypothesis. If this is the case we would see large and significant differences in the school effects estimated based on credits even though these schools may not be any better or worse at improving student achievement.

Table IV.2. School VAM results for grades 6 to 8, by outcome

Outcome	Grades	Adj. R-squared	Students	Schools	SD of School Effects	Mean Standard Error	Statistically Significant Effects	Year-to-Year Correlation
One-Year Model (2009-10)								
PSSA math	6-8	0.75	4520	31	0.21	0.06	0.65	**0.86**
PSSA reading	6-8	0.70	4524	31	0.15	0.07	0.48	**0.76**
PSSA science	8	0.70	1440	29	0.20	0.10	0.41	0.13
PSSA writing	8	0.57	1440	29	0.24	0.12	0.41	**0.50**
CBA math	6-8	0.41	3506	30	0.28	0.11	0.53	n/a
CBA English	6-8	0.52	3480	29	0.15	0.10	0.31	n/a
CBA earth science	6	0.67	1479	27	0.43	0.11	0.67	n/a
CBA life science	7	0.68	1347	29	0.35	0.13	0.38	n/a
CBA physics	8	0.61	1374	29	0.29	0.13	0.31	n/a
CBA US history	8	0.63	1300	27	0.36	0.13	0.52	n/a
SRI	6-8	0.58	3153	29	0.21	0.09	0.52	n/a
Attendance rate	6-8	n/a	4561	31	0.12	0.06	0.35	n/a
Credits earned	6-8	0.86	4548	29	0.97	0.05	0.93	n/a
Two-Year Model (2008-09 to 2009-10)								
PSSA math	6-8	0.74	9321	30	0.19	0.05	0.73	n/a
PSSA reading	6-8	0.70	9316	30	0.18	0.05	0.50	n/a
PSSA science	8	0.69	2978	28	0.11	0.07	0.36	n/a
PSSA writing	8	0.56	2966	28	0.22	0.08	0.46	n/a

Note: The VAMs produce estimates for 24-30 middle and K-8 schools depending on the outcome. CBA scores are not available in all schools. An r-squared value is not calculated in the attendance rate VAM because it uses a different model specification, known as a Tobit model, to adjust for students with perfect attendance (that is, students who are top-coded). Approximately 10 percent of students in the sample have perfect attendance. Bold indicates that year-to-year correlations are statistically different from zero using a 95 percent confidence interval.

Results for Grades 9-10

We now turn to the estimates of high school level VAMs. Table IV.3 summarizes the results. We can distinguish 38% of high schools from average across all grades and outcomes. The year-to-year correlations in school effects are in general high and positive, although they are imprecisely estimated due to the small number of high schools in our sample (9 in most cases). As with the middle-school analyses, a few CBA assessments were entirely missing from a few high schools.

Care should again be taken with the interpretation of the credits outcome, because we cannot be sure that credit-earning is defined consistently across schools. Ensuring a consistent definition will be essential if credits are to be used in future VAMs for high stakes purposes.

Table IV.3. School VAM results for grades 9 to 11, by outcome

Outcome	Adj. R-squared	Number of Students	Number of Schools	SD of School Effects	Mean Standard Error	Number of Statistically Significant Effects	Year-to-Year Correlation
Grade 9 (2009-10)							
CBA Algebra I	0.53	540	9	0.26	0.17	1	0.43
CBA Biology	0.48	794	8	0.33	0.12	4	0.71
CBA Civics	0.56	958	9	0.35	0.10	6	0.54
CBA English I	0.53	903	8	0.12	0.09	1	0.86
SRI (Dec/May)	0.60	962	8	0.10	0.08	1	n/a
PSAT Math*	0.71	1012	9	0.09	0.07	2	0.82
PSAT Reading*	0.66	1038	9	0.10	0.07	2	0.43
Credits	0.45	1335	9	0.61	0.10	6	0.87
Attendance	0.40	1335	9	0.30	0.10	4	0.57
Grade 10 (2009-10)							
CBA Chemistry	0.52	801	8	0.44	0.13	3	n/a
CBA English II	0.57	894	9	0.07	0.08	0	0.61
CBA Geometry	0.64	790	9	0.27	0.10	4	0.69
CBA History	0.42	724	8	0.40	0.11	6	0.09
SRI (Dec/May)	0.55	745	9	0.19	0.11	2	n/a
PSAT Math*	0.73	1023	9	0.03	0.05	0	0.48
PSAT Reading*	0.69	1041	9	0.14	0.07	4	0.90
Credits	0.32	1244	9	0.31	0.09	3	0.72
Attendance	0.56	1244	9	0.11	0.09	4	0.13
Grade 11 (2009-10)							
PSSA Math	0.79	1112	9	0.13	0.06	2	0.72
PSSA Reading	0.76	1111	9	0.10	0.05	1	0.71

Note: A * next to a variable denotes that it is assigned to the teachers from the previous grade and year (e.g. PSAT Math is taken in the beginning of Grade 10, so it is assigned to the grade 9 math teacher from the previous year). Year-to-year correlations of teacher effects are presented for outcomes for which we have more than one year of data.

The high-school results are particularly relevant to the design of the formula for awarding bonuses to teams of Promise Readiness Corps teachers who have moved a cohort of entering 9th graders to successful completion of 10th grade and preparation for 11th grade. The results in Table IV.3, indicating that the models explain a substantial amount of variance for all outcomes, and indicating that VAM estimates for most outcomes are stable from year to year, suggest that all of the outcome measures examined in those two grades could contribute to PRC bonus calculations—although, as noted in Chapter II, most of them should be re-normed annually rather than benchmarked historically, given instability in scales from year to year.

One complication related to the use of CBAs is that they are not used for all levels of coursework in all subjects. In high-school science and perhaps other subjects, students enrolled in advanced courses (known as CAS courses) do not take CBA assessments. We are now undertaking analyses to assess the extent to which students are exempt from CBAs and how much these exemptions vary across schools.

Correlation with PVAAS Results

As part of our reporting of school VAM findings, we compare our findings to results for Pittsburgh schools according to the Pennsylvania Value Added Assessment System (PVAAS). PVAAS data, like several of our VAMs, originate from statistical analyses of longitudinally-linked student PSSA scores. The PVAAS model includes data from more prior years than we do but does not control for socioeconomic or demographic factors at any level. The main PVAAS findings indicate how much academic improvement each cohort made at a school since the prior year. We correlate our VAM results with the mean gain over tested grades at each school relative to the state average, as reported by PVAAS.

Table IV.4 shows the correlation results for school VAMs estimated for grades 4 to 8, which are similar for the most part to findings in our prior report. In each of the last two years, our VAM results have correlated positively with PVAAS estimates. With the exception of the science grade 8 exam in 2009-10, the correlations are large and statistically different from zero. We also find that most of our school VAMs based on PSSA scores in grades 4 to 8 have a moderate level of year-to-year correlation that is comparable to PVAAS. In our prior report, we concluded that PVAAS had a lower degree of year-to-year correlation than our estimates based on analyses just of math and reading scores between 2006-07 and 2008-09. While we see similar results for math and reading using the new year of data as well, PVAAS correlations for writing and science are at least as large as our own. Science in grade 8 also has the lowest degree of intertemporal correlation among our models and is the only set of results where we cannot reject a zero year-to-year relationship.

Table IV.4. Correlation of Mathematica and PVAAS VAM estimates, grades 4 to 8

		Correlation to PVAAS		Intertemporal Correlation	
	Grades Included	2009-10	2008-09	MPR	PVAAS
Math	4-8	**0.62**	**0.58**	**0.68**	0.18
Reading	4-8	**0.81**	**0.56**	**0.59**	0.17
Science	4	**0.92**	**0.92**	**0.36**	**0.57**
Science	8	0.17	**0.79**	0.13	**0.64**
Writing	5	**0.94**	**0.68**	**0.57**	**0.54**
Writing	8	**0.78**	**0.82**	**0.50**	**0.71**

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.

E. Composite Value- Added Estimates

For each school in the district, the analyses described above produce multiple VAM measures, from a minimum of four VAM estimates (for K-5 schools) to eleven or more estimates (for high schools). We have also developed composite measures based on an average of VAM estimates across the individual student test outcomes (excluding the non-test outcomes of attendance and credit accumulation).

Composite measures will be meaningful only if there is a component of school performance that is consistent across outcomes. If there is no relationship between school performance in reading and school performance in math (for example), then composite measures will not provide useful information. In fact, nearly all of the pairwise correlations among value-added estimates for different outcomes are positive, as indicated by the correlation matrices in appendix tables A.2 and A.4. This suggests that a composite could provide a meaningful measurement of a school's value added across a wide range of student outcomes.

We generated composite measures for each school by estimating the VAMs for each assessment together in one model. We implemented this estimation procedure by grouping the data on each assessment together, fully interacting a dummy variable indicating which assessment the data belonged to with all the right-hand side variables, and running one regression. This method produced numerically equivalent estimates for each assessment as those that were generated when the VAMs were all estimated separately. This method also has the advantage of providing estimates of the covariance of the teacher effects across multiple subjects. These covariance terms are necessary for producing correct standard errors when averaging the teacher effects across the outcome measures to produce the composite estimates. Equal weight was given to each assessment when generating the composite measures.

Composite results for schools by grade level are presented in Tables IV.5 and IV.6. In most cases, composites have somewhat greater precision than results for individual subjects, as might be expected. But schools are also clustered closer to each other in the middle of the distribution, so the greater precision of the estimates does not necessarily mean we can distinguish more schools from average.

Table IV.5. Composite school VAM results for grades 4 to 8

Composite	Adj. R-squared	Total Assessments x Grades	Number of Schools	SD of Teacher Effects	Mean Standard Error	Fraction Statistically Significant Effects
Grades 4 to 5						
One-Year Model	0.70	6	39	0.12	0.05	0.44
Two-Year Model	0.69	6	39	0.10	0.04	0.44
Grades 6 to 8						
One-Year Model	0.63	21	29	0.14	0.04	0.48
One-Year Model (PSSA only)	0.71	8	29	0.18	0.05	0.66
Two-Year Model	0.71	8	29	0.19	0.04	0.73

Note: The composite measure is the weighted average of school effects across assessment outcomes in each grade. For example, in grades 4 to 5 there are six total assessments (2x math, 2x reading, 1x science, 1x writing). Composites include PSSA assessment outcomes only, except for the one-year model for grades 6 to 8. That model also includes CBA and SRI assessments.

Table IV.6. Composite School VAM Results for grades 9 to 11

Outcome	Adj. R-squared	Total Assessments x Grades	Number of Schools	SD of School Effects	Mean Standard Error	Number of Statistically Significant Effects
English	0.62	12	9	0.12	0.08	2
Science	0.52	10	9	0.38	0.13	3
Social Studies	0.49	4	9	0.37	0.10	6
Math	0.67	3	9	0.15	0.08	2

Note: The composite measure is the equal weighted average of school effects across all outcomes within the subject. For English there are 12 assessments that make up the composite, 10 for math, 4 for social studies, and 3 for science.

These composite results involve equal weight for each outcome measure, but PPS could choose to develop different weights, if it determines that particular outcome measures are of greater policy importance. We also conducted alternate analyses with weights that varied based on the precision of the results for each test-specific VAM. This method provides greater weight to assessments that produce greater precision in the VAM results. We chose not to present those results, however, for two reasons. First, they did not substantially or consistently increase the proportion of schools that could be reliably distinguished from district-wide average performance. Second, they implied the use of weights that might not be consistent with the district's policy preferences (for example, placing a large weight on PSAT scores).

F. Differentials in School Value- Added by Race

Individual teachers do not serve enough students to allow useful VAM estimates for subsets of student populations, but schools are large enough that it may be possible to examine whether a school's value-added varies for different kinds of students (if the relevant groups of students are large enough). At the request of PPS, we examined the extent to which schools showed significantly different value-added estimates for their white students and African-American students. We found that the school effect estimates across the samples are positively related, with correlations of 0.64 for grades 4-5 and 0.81 for grades 6-8.

For the great majority of Pittsburgh schools, composite value-added estimates did not differ by statistically significant margins for students of different races. As indicated in Table IV.7, in grades 4-5, only two of 28 schools had value-added estimates that differed significantly by race; this number could have shown differences by chance alone. In grades 6-8, five of 22 schools showed value-added estimates that differed significantly by race. Notably, the differences by race were not always in the same direction: some schools showed higher value-added for African-American students, while others showed higher value-added for white students.

Table IV.7. Two- year composite VAMS by race, grades 4 to 8

	Grades 4 to 5		Grades 6 to 8	
	African-American	White	African-American	White
Standard Deviation of School Effects	0.11	0.12	0.22	0.15
Mean Standard Error	0.054	0.059	0.063	0.056
Fraction Distinguishable from Average	0.32	0.41	0.61	0.45
Number of Students	11525	7382	13725	9084
Number of Schools	38	29	33	22
Within School Correlation	0.64		0.81	
Estimates that are Statistically Larger for African-American Students	0 of 28		4 of 22	
Estimates that are Statistically Larger for White Students	2 of 28		1 of 22	

Note: VAM estimates by race are shown only for schools with more than ten students in the analysis sample in each grade range. This is the same threshold we require on all VAMs.

G. Modification for Attrition and Dropout in High Schools

In our August 2010 report, we noted that differential rates of attrition (and particularly dropout) from different high schools around the district could produce biased estimates of value added. We found that attrition rates—i.e., the proportion of entering 9th graders who are excluded from VAMs in subsequent years because they do not remain enrolled in Pittsburgh schools—vary widely among the district's high schools. We also found that, in an analysis of 11th-grade PSSA scores for 2008-09, schools with higher attrition rates tended to have higher estimates for their value-added for the students who remain. This is potentially a problem for the analysis, because it is likely that dropouts were experiencing, on average, lower achievement growth prior to dropping out. In consequence, unadjusted VAM estimates could be biased upward for schools with high dropout rates and downward for schools with low dropout rates.

For this report we have examined the correlation between dropout rates and school effects for each of the Grade 10 and 11 outcome measures in 2009-10, shown in Table IV.8. We continue to observe positive correlations for some outcome measures, particularly the Geometry and History CBAs, and the SRI, but we also observe some negative correlations, including for both Math and Reading Grade 11 PSSA outcomes, and there is no consistent pattern of positive correlations between dropout rates and school performance.

Table IV.8: Correlations of 10th grade outcome measures with school attrition rates

Grade 10 and 11 Outcome Measures	Correlation Between School Outcome Measures and School Attrition Rates
Chemistry CBA	0.35
English 2 CBA	-0.14
Geometry CBA	0.67
History CBA	0.55
Grade 10 May SRI	0.57
PSAT Math	0.21
PSAT Reading	-0.17
Total Credits	-0.51
Attendance	-0.44
Grade 11 PSSA Math	-0.16
Grade 11 PSSA Reading	-0.31

For the grade 10 outcomes where we do observe substantial correlation between the VAM estimates and school attrition rates, Geometry and History CBAs and the SRI, we examined the possibility of using a weighting method to adjust for differential attrition bias. This method is described in the memo we delivered on September 21, and essentially consists of estimating for each student a predicted probability of dropping out; using that predicted dropout probability to create weights for each student so that weighted predicted dropout rates for the students who have not dropped out are the same for each PPS high school; and then running the high school VAM as a weighted regression, giving some students more weight than others in estimating each school's value-added.

Interestingly, using the weighting approach did not consistently reduce the correlations between the school VAM estimates and dropout rates. For the Geometry CBA the correlation fell from 0.67 to 0.35, for the History CBA the correlation rose from 0.55 to 0.72, and for the SRI the correlation fell from 0.57 to 0.51. The fact that the correlations did not fall substantially is likely due to the fact that the weighting adjustment depends on the observable characteristics of students (such as demographics and prior achievement), but students who dropped out are likely different from students who persist in ways that are unobservable. For this reason, and because we do not see evidence of systematic correlations between dropout rates and high school VAM estimates, we do not recommend using the weighting approach to adjust for differential dropout.

We will continue to investigate the issue of differential attrition in subsequent reports. Our briefing on November 30 will include additional information that describes how dropouts differ from other PPS students.

H. Sensitivity Test: Prior- Year Scores vs. Pre- Entry Scores

In our August report, we discussed how prior-year scores in school VAMs typically are not pre-treatment baselines (unlike in teacher VAMs) because most students attended the same school in the prior year. As a result, the baseline scores are not determined outside the model as they are for teachers because schools contribute value added to both baseline and outcome scores. In this section, we conduct a model sensitivity analysis to examine the extent to which using scores that are not pre-treatment scores affects the school VAM estimates. This analysis is motivated by the fact that no pre-treatment scores are available for value-added estimates in elementary grades (because the standardized testing regime begins at grade 3).

Our analysis takes advantage of the fact that scores in grade 5 and in grade 8 are pretreatment baselines for students attending middle and high schools, respectively. We compare models similar to those we presented in August with a specification that instead uses the pre-treatment score regardless of the student's current grade. For example, we would use the grade 5 score regardless of whether the student was enrolled in grade 6, 7, or 8 in 2009-10. Table IV.9 shows results for middle schools and Table IV.10 shows results for high schools.

Table IV.9. Middle- school VAMs using 5th- grade controls versus prior- grade controls

	PSSA Math		PSSA Reading	
	Prior Year Score	5th Grade Score	Prior Year Score	5th Grade Score
1-Year Model (2009-10)				
Standard deviation of school effects	0.19	0.26	0.14	0.19
Mean standard error	0.04	0.05	0.05	0.05
Number (of 11) distinguishable from average	6	6	7	8
Within- school correlation	0.97		0.93	

Table IV.10. High school VAMs using 8th grade- controls versus prior- grade controls

	CBA English II		CBA Geometry		Grade 10 PSAT Math		Grade 10 PSAT Reading	
	8th Grade Controls	Prior-Year Controls	8th Grade Controls	Prior-Year Controls	8th Grade Controls	Prior-Year Controls	8th Grade Controls	Prior-Year Controls
1-Year Model (2009-10)								
Adjusted R-squared	0.57	0.54	0.64	0.55	0.72	0.72	0.69	0.73
Standard deviation of school effects	0.07	0.10	0.27	0.37	0.03	0.03	0.14	0.11
Mean standard error	0.08	0.09	0.10	0.14	0.05	0.06	0.07	0.07
Number (of 9) distinguishable from average	0	1	4	5	0	0	4	1
Within- school correlation	**0.93**		**0.87**		**0.78**		**0.90**	

Note: Statistically distinguishable is defined using a 95 percent confidence interval.

The findings suggest that the models tend to perform similarly under both specifications, although the school estimates are more highly correlated at the middle school level than at the high school level. Specifically, the within-school correlation of estimates is 0.98 for both math and reading in middle schools. It ranges from 0.78 to 0.93, however, for high schools, depending on the outcome. Given the lower correlations across models for at the high school level and our presumption that using pretreatment scores inherently is a superior model design, our base models now use the pretreatment score when one is available. This specification ensures that the model is making attributions to schools controlling only for factors that schools cannot influence. We continue to rely on prior year scores for students in elementary grades because a pre-treatment score is unavailable for these students; the high correlation of VAM results for the two models in grades 6-8 provides reassurance for using a model with prior-year results in elementary grades (although it does not address the challenge of the lack of information on achievement trajectories in grades K-3). For students in grades 6 to 8 at K-8 schools, we use their grade 5 scores regardless of their current grade to maintain comparability with students attending middle schools because both sets of schools are included in school VAMs for grade 6 to 8.

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

V. NEXT STEPS IN VALUE- ADDED MODEL DEVELOPMENT

The findings in this report reinforce the findings of the August report (Lipscomb, Gill, and Booker, 2010) that VAM estimates can provide meaningful information about teacher and school performance in Pittsburgh—and they confirm that VAMs can be usefully applied not only to state accountability tests and nationally normed assessments, but also to locally developed assessments and to non-test measures of student outcomes such as attendance and credit accumulation. We do not find evidence suggesting that any of the locally developed CBAs must be excluded from incorporation in VAMs. The addition of results from CBAs, SRI, PSAT, and science and writing PSSA assessments have allowed far more teachers to be included in value-added analyses, and have dramatically broadened the range of outcomes incorporated in school-level VAM estimates (which now include attendance and credit accumulation as well as many more assessment outcomes).

For most outcomes, VAMs reliably distinguish the highest- and lowest-performing teachers from average with only one year of teaching data, suggesting that the one-year models are diagnostic for some purposes. But we continue to find that the reliability of teacher-level value-added estimates improves substantially when examining two or three years of teaching rather than one year. We therefore recommend using VAM estimates that are based on two or three years of teaching for high-stakes decisions. School-level VAM estimates also tend to have greater ability to detect differences with more than one year of teaching examined, but the improvement in precision is not always as large as it is with teachers.

We find that the inclusion of a statistical control for a student's race/ethnicity makes a measurable difference in the results of many of the VAMs, particularly in high school. Omitting the variable for race/ethnicity therefore could produce value-added estimates that are biased against teachers who serve higher-than-average proportions of African-American students. Given this finding, we recommend continuing to include race/ethnicity as one of the control variables in the VAMs.

In upcoming work, we will continue to refine the VAMs to improve precision and reduce bias. The state of the art in value-added modeling changes rapidly, and we intend to ensure that Pittsburgh's VAMs are updated with refinements developed by our own colleagues and others in the field.

Some of these potential refinements are already on the agenda for upcoming work. As described in Chapter IV, we will examine whether the kindergarten DIBELS assessments can be used to validate a retrospective elementary school VAM that is based on gains from grades 3 to 5. If so, it would increase confidence that VAM estimates using only a subset of elementary grades nevertheless provide a reasonable view of performance across grades K to 5. The exploratory examination of gains achieved between kindergarten DIBELS assessments and third-grade PSSA scores may also open the possibility of extending the elementary-school VAMs to incorporate an analysis of third-grade results that serves as a proxy for gains from kindergarten.

Our examination of differential attrition rates among high schools, for each of the 10th-grade outcomes of interest, suggests that the problem may be less serious than our August findings indicated. Nonetheless, we will conduct further analyses to assess the extent to which the validity of school-level VAM estimates in high school might be threatened by variations in dropout rates across the district.

The analyses of 9th- and 10th-grade outcomes suggest that VAMs using all of the outcomes (with adjustments for attrition) could be appropriately included in a PRC bonus formula. We recommend that the size of the bonus pool be tied to annual changes in district-wide scores on the PSAT, which—unlike the other outcomes relevant to PRC teams—is designed to have consistent scales over time. In future years, the PRC bonus formula might be shifted away from a PSAT anchor to a comparison against statewide results, but this will have to wait until Keystone exams have been implemented in Pittsburgh and across the state.

The generally positive within-school correlations among VAM estimates for different outcomes suggest that a composite measure should provide a useful global indicator of a school's overall value-added across subjects and grades. Preliminary composites of school performance are included in the current report, and future work can adjust those composites to incorporate different weights that are deemed by PPS and PFT to reflect the district's policy preferences regarding the relative importance of different outcomes. We are also prepared to develop composite VAM estimates for the subset of teachers whose students take more than one assessment.

In 2011 we also intend to examine the relationships of teacher VAM results to several other classroom measures, which may include data from teacher working conditions survey, and RISE classroom performance evaluations. We look forward to working with PPS and the PFT in continuing to develop the value-added models to their fullest potential.

REFERENCES

Aaronson, D., Barrow, L., and W. Sander. 2007. Teachers and student achievement in the Chicago public high schools. Journal of Labor Economics 25, (1): 95-135.

Ballou, D. 2005. Value-added assessment: Lessons from Tennessee. In Value Added Models in Education: Theory and Applications, Edited by R. Lissetz. Maple Grove, MN: JAM Press, pp. 272-303.

Booker, T.K. and B. Gill. 2010, September. Alternative method for addressing differential attrition in high-school value added. Memo to Pittsburgh Public Schools VAM Development Team.

Engberg, J., and B.P. Gill. 2006. Estimating graduation and dropout rates with longitudinal data: A case study in the Pittsburgh Public Schools. Santa Monica, CA: RAND.

Goldhaber, D. and M. Hansen. 2008. Assessing the potential of using value added-estimates of teacher job performance for making tenure decisions. Center on Reinventing Public Education, working paper #2009-2.

Hanushek, E. A., and S. G. Rivkin. 2008. Do disadvantaged urban schools lose their best teachers? The Urban Institute, National Center for Analysis of Longitudinal Data in Education Research.

Hanushek, E. A., and S. G. Rivkin. 2010. Using value-added measures of teacher quality. The Urban Institute, National Center for Analysis of Longitudinal Data in Education Research.

Jones, M. P. 1996. Indicator and stratification methods for missing explanatory variables in multiple linear regression. Journal of the American Statistical Association 91, (433): 222-230

Kane, T. J. and D. O. Staiger. 2002. The promises and pitfalls of using imprecise school accountability measures. Journal of Economic Perspectives 16, (4): 91-114.

Kane, T. J. and D. Staiger. 2008. Estimating teacher impacts on student achievement: An experimental evaluation. NBER Working Paper.

Koedel, C. and J. Betts. 2010. “Value Added to what? how a Ceiling in the Testing Instrument Influences Value-Added Estimation.” Education Finance and Policy, vol. 5, no. 1, pp. 54-81..

Krueger, A. B. 1999. Experimental estimates of education production functions. Quarterly Journal of Economics 114, (2): 497-532.

Lipscomb, S., B. Gill, and T.K. Booker. 2010, August. Estimating teacher and school effectiveness in Pittsburgh: value-added modeling and results. Draft report. Cambridge, MA: Mathematica Policy Research.

McCaffrey, D. F., Sass, T. R., Lockwood, J. R., and K. Mihaly. 2009. The intertemporal variability of teacher effect estimates. Education Finance and Policy 4, (4): 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. 1983. Parametric empirical Bayes inference: Theory and applications. Journal of American Statistical Association 78, (381): 47-55.

Potamites, L., Booker, K., Chaplin, D., and Isenberg, E. 2009. 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.

Potamites, L., Chaplin, D., Isenberg, E., and Booker, K. 2009. Measuring school effectiveness in Memphis – Year 2. Final report submitted to New Leaders for New Schools. Washington, DC: Mathematica Policy Research.

Rothstein, J. 2010. Teacher quality in educational production: Tracking, decay, and student achievement. Quarterly Journal of Economics 125, (1): 175-214.

Schochet, P. Z. and Chiang, H. S. 2010. Error rates in measuring teacher and school performance based on student test score gains. Washington, DC: U.S. Department of Education.

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

APPENDIX A

SUPPLEMENTARY TABLES AND FIGURES

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

Table A.1. Within-teacher VAM correlations across outcome measures, grades 4-5

	PSSA Math	PSSA Reading	PSSA Science
One-year model (2009-10)			
PSSA Math	1		
PSSA Reading	**0.56**	1	
PSSA Science	0.19	0.29	1
PSSA Writing	0.32	**0.51**	

Note: Bold indicates statistically significant correlations using a 95 percent confidence interval.

Table A.2. Within- school VAM correlations across outcome measures, grades 4- 5

	PSSA Math	PSSA Reading	PSSA Science	PSSA Writing
One-year model (2009-10)				
PSSA Math	1			
PSSA Reading	0.3	1		
PSSA Science	0.26	0.18	1	
PSSA Writing	0.21	**0.56**	0.31	1
Attendance rate	0.13	0.2	0.01	0.29

Note: Bold indicates statistically significant correlations using a 95 percent confidence interval.

Table A.3. Within- teacher VAM correlations across outcome measures, grades 6- 8

	PSSA Math	PSSA Reading	PSSA Science	PSSA Writing	CBA math	CBA English	CBA earth science	CBA life science	CBA physics	CBA U.S. history	SRI
One-year Model (2009-10)											
PSSA Math	1										
PSSA Reading		1									
PSSA Science			1								
PSSA Writing		**0.61**		1							
CBA math					1						
CBA English		**0.31**		0.25		1					
CBA earth science			0.24		-0.18		1				
CBA life science			0.05				**0.8**	1			
CBA physics			0.15				**0.85**	**0.49**	1		
CBA U.S. history										1	
SRI		**0.2526**		**0.37**		**0.28**					1

Note: Bold indicates statistically significant correlations using a 95 percent confidence interval.

Table A.4. Within- school VAM correlations across outcome measures, grades 6- 8

	PSSA Math	PSSA Reading	PSSA Science	PSSA Writing	CBA math	CBA English	CBA earth science	CBA life science	CBA physics	CBA U.S. history	SRI	Attendance rate
One-year Model (2009-10)												
PSSA Math	1.00											
PSSA Reading	**0.69**	1.00										
PSSA Science	**0.60**	**0.57**	1.00									
PSSA Writing	**0.53**	**0.65**	0.63	1.00								
CBA math	**0.55**	0.26	**0.49**	0.12	1.00							
CBA English	**0.70**	**0.71**	**0.48**	**0.58**	**0.51**	1.00						
CBA earth science	0.03	0.00	-0.09	-0.05	0.04	0.02	1.00					
CBA life science	0.12	-0.07	-0.06	-0.24	0.18	-0.19	0.23	1.00				
CBA physics	0.05	0.04	0.30	0.04	0.16	0.09	**0.54**	0.25	1.00			
CBA U.S. history	0.09	0.35	0.33	0.31	-0.06	0.03	0.03	-0.21	-0.01	1.00		
SRI	0.36	**0.42**	0.34	**0.42**	0.30	**0.47**	0.15	-0.13	0.11	0.38	1.00	
Attendance rate	-0.36	-0.22	-0.06	-0.22	-0.10	-0.27	-0.14	0.11	0.02	-0.22	-0.33	1.00
Credits earned	0.20	0.32	0.32	0.17	**0.38**	**0.47**	0.12	-0.17	0.12	-0.02	0.26	0.19

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