Value-Added Estimates for Phase 1 of the Pennsylvania Teacher and Principal Evaluation Pilot

Full Report

April 5, 2012

Stephen Lipscomb
Hanley Chiang
Brian Gill

MATHEMATICA
Policy Research

All statistics are calculated by Mathematica unless stated otherwise

Mathematica Reference Number:
06815.300

Submitted to:
Team Pennsylvania Foundation
100 Pine Street, 9th Floor
Harrisburg, PA 17101
Project Officer: Jennifer Cleghorn

Office of the Deputy Secretary of Elementary and Secondary Education
Pennsylvania Department of Education
333 Market Street
Harrisburg, PA 17126-0333
Project Officer: Carolyn C. Dumaresq

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

Value-Added Estimates for Phase 1 of the Pennsylvania Teacher and Principal Evaluation Pilot

Full Report

April 5, 2012

Stephen Lipscomb
Hanley Chiang
Brian Gill

All statistics are calculated by Mathematica unless stated otherwise

ACKNOWLEDGMENTS

This report would not be possible without the support of the Team Pennsylvania Foundation, the Pennsylvania Department of Education, the Bill & Melinda Gates Foundation, and members of the Stakeholder Steering Committee who volunteered their time during Phase 1 of this pilot study. We are especially grateful to Carolyn Dumaresq, Jennifer Cleghorn, Matt Zieger, Pat Hardy, Rich Maraschiello, Sharon Brumbaugh, and Theresa Barnaby for their project leadership, to Harris Zwerling for his insightful comments and suggestions, and to Dave Ream, Thomas Gillin, and Thresa Stafford for providing us with the statewide data used in this report. We also thank Bob Hughes, Suzanne Lane, and the teachers, principals, and superintendents from the four pilot school districts for their important contributions during the pilot toward the improvement of the evaluation system.

We would like to recognize several staff at Mathematica Policy Research who made substantive contributions to this report. Clare Wolfendale provided outstanding research support in estimating the value-added models and populating tables of results. We are also grateful to Serge Lukashanets, Xiaofan Sun, and Alena Davidoff-Gore for their contributions to organizing the data, to Duncan Chaplin for his comments and suggestions on an earlier version, to John Kennedy for his editorial assistance, and to Autumn Parker and Eileen Curley for their production support.

Any opinions expressed herein are those of the authors and do not necessarily represent the views of the Team Pennsylvania Foundation or the Pennsylvania Department of Education.

THIS PAGE LEFT BLANK FOR DOUBLE- SIDED PRINTING

CONTENTS

V *(continued)*

TABLES

FIGURES

THIS PAGE LEFT BLANK FOR DOUBLE- SIDED PRINTING

EXECUTIVE SUMMARY

The Commonwealth of Pennsylvania plans to develop a new statewide evaluation system for teachers and principals in its public schools by school year 2013–2014. To inform the development of this evaluation system, the Team Pennsylvania Foundation (Team PA) undertook the first phase of the Pennsylvania Teacher and Principal Evaluation Pilot—henceforth referred to as Phase 1—in 2010 and 2011 in collaboration with a broad stakeholder group that included leaders from the Pennsylvania Department of Education (PDE), the Pennsylvania State Education Association (PSEA), school districts, and the business community. The purpose of Phase 1 was to develop and implement a pilot set of performance measures to obtain lessons for improving the use of classroom observations and student data in evaluating teacher and principal performance. None of the results from Phase 1 had a bearing on actual evaluations or personnel decisions for any teacher or principal.

Phase 1 proceeded along two tracks. In the first track, observation-based rubrics for evaluating teacher and principal effectiveness were implemented on a trial basis in the Allentown, Cornell, and Mohawk Area school districts, and in Northwest Tri-County Intermediate Unit 5 (collectively referred to as Phase 1 pilot districts). Based on these rubrics, a set of preselected principals and teachers from the pilot districts were rated by their supervising superintendents and principals, respectively, in spring 2011. Lane and Horner (2011) discuss the results of this track.

This report presents findings for the second track of Phase 1. In this track, Mathematica Policy Research used student-level data to develop value-added models (VAMs) for estimating teacher and principal effectiveness. VAMs estimate the effects of educators on student achievement growth. VAMs belong to the class of models that are generally referred to as student growth models, but a VAM estimate is not a measure of student growth; rather, it is an estimate of an educator's or a school's *contribution* to student growth. VAMs can be appropriate for use in teacher or principal evaluations because they produce information about educator effectiveness. Other indicators like student proficiency rates and descriptive measures of student growth might be appropriate as targets for school accountability purposes, but they should not be viewed as indicating what a teacher or school has contributed to student learning.

After calculating these effectiveness estimates, Mathematica then examined whether Phase 1 teachers with higher classroom observation scores on specific professional practices covered by the pilot rubric had greater impacts on student achievement as measured by value-added.

Specifically, we address the following three primary research questions in this report:

1. How can VAMs be used to characterize the effectiveness of teachers at raising achievement according to multiple outcome measures?
2. Do specific teacher practices relate to larger contributions to student learning among Phase 1 teachers?
3. How can principals' contributions to student learning be measured?

The U.S. Department of Education's Race to the Top initiative is a prominent example of the interest among federal, state, and local policymakers in measuring educator effectiveness based on performance, and VAMs have been a focal point in these debates. **In a VAM, the actual level of achievement demonstrated by an educator's students is compared to the level that would be**

predicted after accounting for students' own prior achievement histories and factors such as the characteristics of their family backgrounds and peers. The differential amount (above or below zero) is averaged across students taught by each educator and attributed to educators as their contribution to achievement. VAMs measure relative teacher performance based on the assessments that are used in the models. The value of VAMs depends in significant part on the validity of the underlying student assessments in capturing what students ought to be learning and the capacity of the tests to allow VAMs to capture meaningful distinctions in achievement. In principle, VAMs can be applied to any quantifiable measure of student outcomes. **As a measure of educator quality, a VAM's fairness depends on whether the method successfully removes influences outside an educator's control.** VAMs do not indicate what level of value-added Pennsylvania should view as adequate in terms of an external standard for specifying whether students are learning "enough." VAMs also do not indicate whether the assessments on which they are based capture the skills that students ought to be learning in the classroom.

We find that VAMs based on multiple outcome measures can be informative tools for identifying highly effective and highly ineffective teachers and schools. However, larger samples of teachers than were available in Phase 1 are needed to ascertain the relationships between instructional practices and teachers' impacts on student outcomes. **VAMs also face limitations in their ability to distinguish educators' true effects—especially the effects of principals—from factors beyond their control, and it is important to take these limitations into account when applying VAMs to a real, large-scale evaluation system. Subsequent phases of the pilot will require additional work to further explore and address these limitations.**

The following sections describe the main findings from the analyses and how these findings should inform the next phase of the pilot.

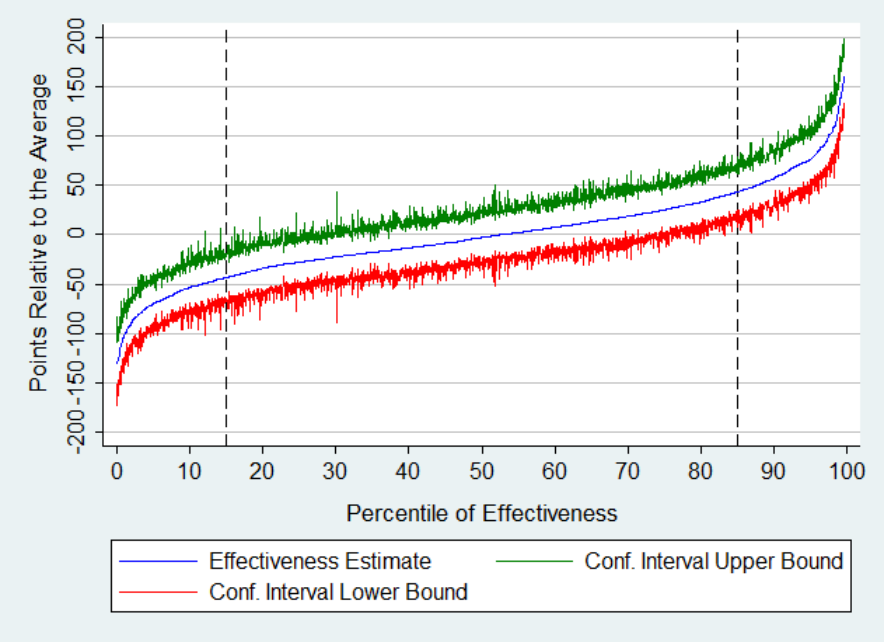
Using VAMs to Estimate Teacher Effectiveness

Teacher contributions to student achievement vary substantially across Pennsylvania.

The size of teachers' effects on students' Pennsylvania System of School Assessment (PSSA) scores varies substantially across the state in all PSSA subjects in grades 4 through 8. In Figure 1, we provide an example of a statewide distribution of teacher effectiveness by depicting it for 5th-grade math teachers. The blue curve indicates the value-added of individual teachers, who are rank ordered along the horizontal axis based on the estimated size of their contribution to 5th-grade math PSSA achievement. Value-added is expressed along the vertical axis in terms of additional PSSA scale points relative to the teacher in the middle of the distribution.[1] For instance, switching from the 15th to the 85th percentile teacher would enable a 5th-grade student who originally scored better than half of all students in the state on the math PSSA to improve by 87 scaled score points and end up scoring better than 65 percent of all students.

[1] Value-added is calculated in terms of z-scores (see Appendix C). We convert z-score units to PSSA scaled scores for illustrative purposes in reporting results.

Figure 1. Distribution of Teacher Effectiveness for 5th- Grade Math PSSA Scores



Source: Mathematica calculations based on Pennsylvania data. The sample includes 2,836 teachers who taught 5th-grade students in each year between 2008-2009 and 2010-2011.

Note: See Figure III.1 for more information. Dashed lines demarcate the 15th and 85th percentiles.

Value-added data has an advantage over most other types of effectiveness information because it can indicate whether the effectiveness of two educators is statistically different. That is, a VAM can indicate with a high degree of confidence whether the actual effectiveness of teachers with low or high VAM estimates is likely to differ from the effectiveness of a teacher in the middle of the distribution. This is the purpose for the intervals around the blue curve in Figure 1, which are called confidence intervals. Statistically speaking, teachers with confidence intervals that are entirely above or below the value-added of the 50th percentile teacher are said to be performing differently from (that is, either above or below) average. Such intervals are characteristic of nearly all of Pennsylvania's 5th-grade math teachers below the 15th percentile and above the 85th percentile. Intervals for teachers closer to the 50th percentile include zero, meaning that their contribution to student achievement growth is typical for 5th-grade math teachers in the state. In short, VAMs have the ability to delineate groups of teachers that differ in their performance estimates to an extent that could not have arisen by chance errors in estimation. Other types of evaluation data like classroom observation data can place teachers into performance categories but cannot indicate whether the performance of teachers across those categories is statistically different unless a confidence interval is reported.

Incorporating multiple student cohorts improves the reliability of effectiveness estimates.

A key design element for a VAM is the number of student cohorts—the full roster of students taught by a teacher in each single year—whose outcomes will factor into a teacher's effectiveness estimate. Outcomes for multiple student cohorts carry potential information on a teacher's contribution. Incorporating students from multiple cohorts in a VAM thus facilitates measuring a teacher's effectiveness with greater statistical reliability. As shown in Table 1, a greater share of the effectiveness estimates can be statistically distinguished from average effectiveness in teacher VAMs

that use three cohorts than in those that use one cohort. Greater reliability is a highly desirable feature for teacher evaluation measures, but the decision to incorporate data from multiple student cohorts comes with tradeoffs. First, with more cohorts, a teacher's effectiveness estimate will be less reflective of the teacher's most recent performance. Second, fewer teachers will have estimates reported that are based on the full number of cohorts used in the VAM, although estimates can be calculated for all teachers based on the number of cohorts available to each.

Table 1. Number of Teachers with Effectiveness Estimates Reported Based on the Number of Cohorts in the VAM and Share of Reported Estimates that Are Statistically Different from the Average

	Number of Teachers with Estimates Reported		Percentage of Reported Estimates that Are Statistically Different from Average	
Outcome	1-Cohort Model	3-Cohort Model	1-Cohort Model	3-Cohort Model
Math PSSA, Grade 5	4,103	2,836	36.5	52.0
Reading PSSA, Grade 8	1,916	1,717	22.3	30.5
Science PSSA, Grade 4	4,187	2,854	27.7	49.8

Source: Mathematica calculations based on Pennsylvania data.

Note: See Table III.3 for more information.

There is more variation in teacher effectiveness within schools than across schools.

About 62 percent of the variation in estimated teacher effectiveness in Pennsylvania is observed within individual schools. This implies that across the state there are plenty of effective teachers in low-performing schools and ineffective teachers in high-performing schools. This finding supports the conclusion that the most important factors to include in a VAM for isolating a teacher's contribution are those that vary within schools.

The remaining 38 percent of the variation is explained by differences in schoolwide average value-added, and this part of the variation poses an analytic dilemma. Average value-added varies from school to school, but is this variation simply the result of the sorting of effective and ineffective teachers, or are the schools affecting their teachers' value-added? The data do not allow us to determine whether the 38 percent of teacher value-added is attributable to the teachers themselves (that is, because good teachers tend to land in the same schools with other good teachers) or to factors at the school that are outside the teachers' control like resource distribution or the quality of the principal. If all of the 38 percent is related to schoolwide factors rather than to teachers, then the VAM should include a control for each individual school—thereby making teachers responsible only for the difference between their own value-added and the average value-added in their schools. This would involve the implicit assumption that average teacher quality is essentially equal in every school across the state, which seems implausible. It could also produce conflicting incentives for teachers. Good teachers in good schools could improve their value-added by moving to low-performing schools. However, absent any movement across schools, teachers could improve their value-added only by performing better than their colleagues down the hall.

Another approach would be to control explicitly for observable school characteristics in the VAM, but there are analytic challenges in determining how to ensure that these adjustments do not absorb true differences in teacher effectiveness across schools. Exploring potential ways of adjusting for school characteristics deserves further attention in Phase 2. For now, the teacher VAMs we use

do not make any school-level adjustments, meaning that teachers are compared with all other teachers (of the same grade and subject) throughout the state and all unmeasured school-level factors relevant to value-added are assumed to be the same across schools.

VAMs based on non-PSSA outcomes have varying degrees of statistical reliability.

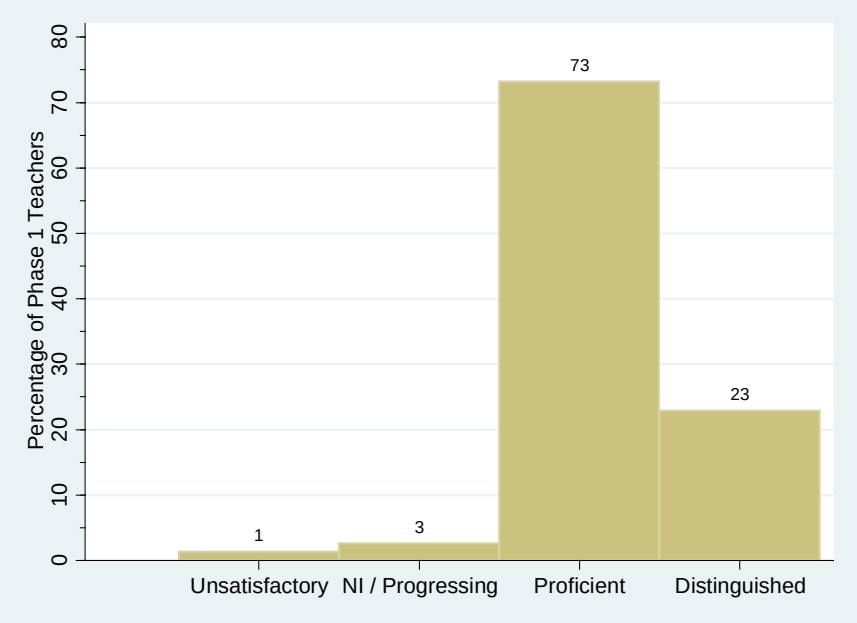
We estimated VAMs based on several non-PSSA outcomes and found notable differences in the ability of the VAMs to make reliable distinctions among teachers. For example, whereas 38 percent of teacher effectiveness estimates can be statistically distinguished from the average on the basis of a 1st-grade curriculum-based writing assessment in Allentown, only 18 percent can be distinguished from the average based on a 2nd-grade measure of early literacy skills. VAMs with greater reliability are likely to be better predictors of teacher abilities in the future as measured by value-added. Therefore, the differences in reliability could be factors in determining what weights PDE would like to place on different types of effectiveness estimates in the evaluation system. As indicated earlier, PDE will want to consider the degree to which the assessments capture the full curriculum in determining how much weight to give particular measures as well.

Teacher Value- Added and the Pilot Observation Rubric

Principals rated nearly all Phase 1 teachers as proficient or distinguished.

In 2011, PDE found that, under the existing evaluation system, principals rated more than 99 percent of teachers across the state as satisfactory. Identifying the bottom 1 percent could be very useful for tenure or other personnel decisions, but the lack of variation in the other 99 percent was a cause for concern. During Phase 1, principals implemented a pilot rubric for teacher observations based on the Framework for Teaching by Charlotte Danielson that included three categories above an unsatisfactory rating. The pilot implementation produced nearly the same result in terms of the percentage of teachers at the low end of the evaluation scale. Specifically, 1 percent of all Phase 1 teachers were rated as unsatisfactory, 3 percent were rated as needing improvement—called progressing for new teachers—and 96 percent were rated as proficient or distinguished (Figure 2). Observation data in Phase 1 were obtained by 30 evaluators for 153 total teachers in the four participating school districts.

Figure 2. Distribution of Final Rating Scores for Phase 1 Teachers



Note: See Figure IV.1 for more information.

NI = Needs Improvement.

In contrast to the existing system, teacher value-added data can reliably distinguish more teachers from average, at the top of the scale as well as the bottom. Depending on the outcome, number of students, and number of years of available teaching data, each low or high performance group usually includes between 15 to 25 percent of teachers, with 50 to 70 percent of teachers in the middle (i.e., not distinguishable from average).

The distribution of observation scores includes a far greater proportion of teachers at the higher performance levels than would be expected based on a normal bell curve. There are at least two reasons why the distribution of scores could be skewed. First, Phase 1 included a very small number of teachers, and those who were sampled were selected by their principals based on no previous evidence of unsatisfactory performance. The scores of these teachers thus may not be representative of scores that would be obtained by teachers across Pennsylvania. However, we do not see evidence to support this possibility, at least based on broad comparisons of the characteristics of pilot teachers and other educators in Pennsylvania. Second, there is some evidence that principals were unwilling to use all available categories to differentiate teachers because evaluators in one pilot district gave all of their teachers exactly the same rating on all components of the observation rubric.

There are no statistically significant relationships between teachers' observation scores and their value-added scores in the Phase 1 data.

Using statistical models, we tested the relationships between teachers' estimated contributions to student learning and their observation scores for the 81 teachers with observation scores and value-added data. The models compared the VAM score for individual teachers with their rubric ratings on each component and overall across components. The analyses sought to measure the predicted increase in teacher contributions to student learning from a one-level increase (for example, from proficient to distinguished) on any component of the observation rubric. Due to the small size of the pilot and the compressed distribution of observation scores, none of the

relationships we estimated are statistically significant. This could change in Phase 2 when a much larger number of teachers will be involved; the research literature includes several studies that indicate that teachers who have higher scores on observational rubrics make larger contributions to student achievement than teachers with lower scores. But if principals are unwilling or unable to differentiate among teachers in their observations, and if 96 percent of teachers again have ratings in the top two categories, we might again find no statistical relationship to value-added estimates. The value of a four-category rubric for professional practice depends on the willingness of the raters to use all of the categories.

Using VAMs to Estimate Principal Effectiveness

The best available method for distinguishing principals' effects on student outcomes from the effects of other school-specific factors can be applied only to a limited number of principals and therefore is not applicable to a real evaluation system.

A key analytic challenge of any principal VAM is to disentangle principals' true contributions to student outcomes from the influence of other school-level factors. A natural starting point for estimating principal effectiveness is to estimate the effectiveness of the principal's school. The complication is that a school's effectiveness can also reflect other school-specific characteristics and circumstances beyond the principal's control, most notably including the preexisting abilities of the school's teachers. Teachers have direct instructional contact with students, but principals can influence student achievement only indirectly.

The best available VAM for isolating pure principal effects, which we call the principal transitions model, calculates how the same school's value-added differs under the leadership of different principals. Thus, it measures how effective a principal is relative to the other principals who have served at the same school. This approach has the benefit of controlling for all school-specific factors beyond principals' control that remain constant over time.

From the statewide data, we identified two major reasons why this method cannot be applied to real-world evaluations of principals. First, it can generate effectiveness estimates for only a limited group of principals—those principals from schools that have undergone leadership transitions. In the statewide data, only a minority of schools underwent leadership transitions over a three-year period. Second, the principal transitions model also limits the ways in which principals can be compared on their performance. Comparisons can be made only within small networks of schools connected by a series of principal transfers. We found that most such networks encompassed only one or two schools, implying that this model measures a principal's effectiveness relative to a very limited comparison group.

VAMs for measuring school effectiveness provide informative but imperfect measures of principals' contributions to student learning.

An alternative model, which is applicable to real evaluations, gives each principal a value-added score based on the average effectiveness of the principal's school(s) during the analysis period. Although this model generates estimates for principals even if they have served in multiple schools, we call it a school VAM to emphasize the fact that it bundles together principals' true contributions with the effects of other school-level factors.

We assessed the degree to which effectiveness estimates from the school VAM deviate from pure principal effects. Estimates from the principal transitions model served as benchmarks with which estimates from the school VAM (for the same principals) were compared. We found a moderate degree of consistency between the effectiveness rankings produced by the two models. About half of principals are placed into identical quartiles of performance by the two models. However, a noticeable minority of principals receive a ranking from the school VAM that differs substantially from their ranking from the transitions model.

School VAM estimates actually capture the contributions of entire schools, including some factors beyond principals' control. Nevertheless, given the moderate consistency of these estimates with those from the transitions model, some of the variation in these estimates among principals is likely to capture true differences in principal quality.

VAMs can generally distinguish among schools with respect to impacts on student assessment scores.

There are sizable differences among schools in VAM estimates. By switching from the 15th to the 85th percentile school, a 5th-grade student who originally scored better than half of all students in the state on the math PSSA would improve by 83 scaled score points and end up scoring better than two-thirds of all students. Moreover, performance differences among schools are estimated with greater statistical reliability than those among teachers due to larger student samples per school. In three-cohort models, typically at least two-thirds of schools can be statistically distinguished from the average based on math PSSA outcomes, and at least half can be distinguished from the average based on reading PSSA outcomes. These are, of course, differences in the total value-added of each principal's school(s). The proportion of the variation that is attributable to the principals themselves (versus other school characteristics that might be outside principals' control) is unknown.

Schools differ in their effectiveness at keeping students enrolled in high school.

We examined VAMs based on a nontest outcome called holding power, or the extent to which high-school students stay enrolled in a Pennsylvania school the following year; this might be viewed as a proxy for a school's effectiveness in preventing dropout. Impacts on holding power differ greatly between the worst-performing schools and all other schools in the state. For instance, the bottom 6 percent of schools lower their 9th graders' probability of enrolling in the following year by more than 30 percentage points relative to the average school. It is worth noting that the validity of these estimates depends on the assumption that the statewide data system has complete records on student enrollment. These estimates also do not include 12th graders, so they do not capture actual graduation outcomes. The data to study 12th graders will not be available until Phase 2 at the soonest. Despite these caveats, school effectiveness estimates for holding power appear to be an informative tool for identifying high schools that perform poorly in keeping their students enrolled in Pennsylvania's public schools.

Looking Ahead to Subsequent Pilot Phases

We offer several recommendations that relate broadly to strategies for sampling educators from the pilot districts and steps for refining and improving the performance measures. With regard to sampling, we recommend oversampling educators for whom we can generate value-added estimates with the greatest validity and relevance to the future evaluation model. In particular, because a future statewide evaluation model will almost certainly include the PSSA, we recommend including a

substantial number of math and English language arts teachers from grades 4 through 8 and science teachers in grades 4 and 8. We also recommend oversampling middle school principals when a new principal evaluation instrument is developed. Given that all middle school grades are tested by the PSSA, value-added scores and rubric scores will cover exactly the same grades for this set of principals. Additionally, teachers and principals should be recruited for the pilot to provide for more variation in the observation measure. Focusing on a limited range of performance inhibits the pilot's ability to differentiate between the practices of more and less effective educators.

Several steps can also be taken to improve the performance measures from the VAMs and the observational rubric. First, the assessment properties of the student outcomes—especially district-administered assessments—and the observational rubrics should be evaluated. This includes assessing interobserver agreement, or the rate at which different observers independently agree on a teacher's observation rating, and observer drift, or the tendency of two raters to agree with each other more frequently over time. Second, the quality of data linkages in Pennsylvania's student data should continue to be evaluated. Third, additional nonassessment outcomes for principal evaluations should be examined, such as by developing value-added models based on 12th-grade graduation outcomes. Fourth, the pilot should continue its progress toward identifying how different types of effectiveness data will be integrated in the overall evaluation model. We look forward to continuing our work on these efforts in Phase 2.

THIS PAGE LEFT BLANK FOR DOUBLE- SIDED PRINTING

I. INTRODUCTION

A. The Pennsylvania Teacher and Principal Evaluation Pilot

The Team Pennsylvania Foundation (Team PA) recognizes that the evaluation of teachers and principals is a critical foundation for the education reforms envisioned by the state's leaders. To develop an evaluation system that is accurate and fair, between 2010 and 2011 Team PA undertook the first phase of the Pennsylvania Teacher and Principal Evaluation Pilot (referred to as Phase 1) from which lessons learned will inform the development of a full, statewide evaluation system by 2013–2014.[2] Phase 1 proceeded along two tracks in collaboration with a broad stakeholder group that included representatives from the Pennsylvania Department of Education (PDE), the Pennsylvania State Education Association (PSEA), school districts, and the business community. The tracks were designed to pilot the development and implementation of measures that would improve the use of both classroom observations and student data in evaluating teacher and principal performance. None of the results from Phase 1 had a bearing on actual evaluations or personnel decisions for any teacher or principal.

In the first track, steering committee subgroups initially developed new observation-based rubrics for evaluating teachers and principals during fall 2010. In January 2011, principals and superintendents from Allentown, Cornell, and Mohawk Area school districts, and from Northwest Tri-County Intermediate Unit 5 (collectively, the Phase 1 pilot districts), were trained in the new protocols.[3] These school and district leaders then implemented the new rubrics on a trial basis during the spring semester in their own districts to 153 preselected teachers and 30 preselected principals, respectively. Lane and Horner (2011) documented the process, progress, and lessons learned from the trial implementation in preparation for Phase 2, which will scale up the pilot to include educators from approximately 100 school districts starting in 2012.

In this report, we present findings for the second track of Phase 1. The second track involved using student data to develop value-added models (VAMs) for measuring teacher and principal contributions to student learning, and then examining professional practices that are positively associated with VAM estimates. We used data from the entire state of Pennsylvania for most analyses; we used data for districts covered in the first track of Phase 1 for other analyses.

A VAM is a statistical model that predicts students' levels of achievement based on students' own achievement histories and other characteristics. The difference between students' actual and predicted achievement (above or below zero) is averaged and attributed to their teachers or schools as a measure of the educators' contributions to student learning. Mathematica developed the VAMs for Team PA in Phase 1 and conducted analyses to address the three primary research questions for this report:

[2] Phase 1 was supported through a grant from the Bill & Melinda Gates Foundation to Team PA.

[3] The National Institute for School Leadership developed and administered the training.

1. How can VAMs be used to characterize the effectiveness of teachers at raising achievement according to multiple outcome measures?
2. Do specific teacher practices relate to larger contributions to student learning among Phase 1 teachers?
3. How can principals' contributions to student learning be measured?

Several of the analyses done in this Phase 1 report are based on small samples of teachers, principals, and schools. Findings from these analyses *should be viewed as providing suggestive evidence that merits further attention in Phase 2 of the pilot.* For instance, the study's second research question relies on data collected once on 153 preselected teachers who teach in four school districts that are not representative of the state in terms of the characteristics of their students.[4] The much larger Phase 2 pilot will provide results that are more precise and more representative of Pennsylvania teachers. To prepare for Phase 2, we invite feedback on how to refine the VAMs in this report to best reflect policy goals for the statewide model evaluation system.[5]

B. Description of Value- Added Models

A well-constructed VAM uses the prior achievement histories of individual students to produce valid estimates of what educators contribute to achievement, regardless of the starting points of their students. VAM estimates overcome a main deficiency of most levels-based measures, such as the rate of student proficiency, which penalize teachers and schools that serve historically low-performing students. By accounting for other observable background characteristics—such as socioeconomic or disability status—of the students assigned to each teacher or principal, VAMs can also overcome a main deficiency of simple growth-in-achievement models that penalize teachers and principals who serve at-risk or hard-to-teach students. Despite these advantages, VAMs—like all measures of performance—are imperfect measures. We recommend basing teacher and principal policy decisions, when possible, on multiple types of information that are combined in an optimal way to ascertain an individual's effectiveness as accurately and completely as possible.

1. Conceptual Framework

The process of estimating a value-added model includes two conceptual steps. In the first step, the VAM makes a prediction about an outcome of interest, typically a student's assessment score in a subject. This prediction is based on factors that include students' own achievement histories and usually other background characteristics about students and their peers. The prediction is derived using data on the performance of other students, either across Pennsylvania or the pilot districts, and represents what we expect a student to achieve if served by the teacher or school in the middle of the effectiveness distribution. It is derived from outcomes achieved by the other students in the same year; the word *prediction* does not mean that a VAM can project a student's future achievement. In the second step, the VAM compares students' actual outcomes with their predicted outcomes.

[4] Analyses of principal practices were not conducted in Phase 1 because the observation rubric for principals will undergo substantial changes during Phase 2 and because the principal pilot was so small that meaningful analyses would not have been feasible.

[5] The overall structure of the teacher and principal evaluation system is under development by PDE. Mathematica is not aware of any plans to include the VAMs developed specifically for this report in the evaluation system.

The VAM score for a teacher or school is the difference between actual performance and the predictions averaged across all students taught by a given educator.

Thus, a value-added model addresses the following central question: *To what extent does the actual level of achievement demonstrated by an educator's students exceed (or fall short of) the level that would have been expected for students with similar achievement histories and similar background characteristics if they had been taught by the educator in the middle of the effectiveness distribution?* A VAM does not measure student achievement growth. It instead seeks to produce something approaching a causal inference about the individual contributions of educators to the learning of students under their charge. Given the available data, VAMs arguably represent the best method for estimating educators' contributions to student learning as measured by assessment scores, but there are likely to be at least some factors that limit the accuracy and validity of these estimates.

Rothstein (2010) concluded that teacher effectiveness measures according to most VAMs lack validity because some teachers are more likely than their colleagues to be assigned students with particularly high or low gains in the previous grade. Fortunately, the degree of bias from this kind of sorting of students might not be large. Kane and Staiger (2008) compared teacher VAM estimates in Los Angeles under a typical situation in which principals assigned students to teachers to VAM estimates in the following year when principals randomly assigned teaching assignments—thereby eliminating the possibility of bias due to the sorting of students. They found that a higher VAM score before random assignment was a positive and significant predictor of achievement differences when classrooms were assigned randomly. In addition, Koedel and Betts (2011) found that the sorting bias identified by Rothstein can be reduced to statistical insignificance by including students from multiple cohorts in teacher VAM estimates, rather than just one cohort as in the Rothstein study. Goldhaber and Chaplin (2011) found that even without using multiple cohorts of students, the bias identified by the specification tests Rothstein uses might be very small.

Another reason to be cautious about interpreting a VAM estimate is that VAMs likely cannot control for all of the relevant factors needed to distinguish completely the teacher's or the school's contribution from every other factor affecting the performance of students. A VAM can control only for those factors that are observable in the data. If there are other student, peer, and school characteristics that influence student performance and that are not captured in the VAMs, they can artificially inflate the VAM estimates for some teachers and deflate the estimates for others.

A final consideration for interpreting the performance measures produced by VAM methods is that VAMs do not measure student achievement growth in absolute terms. They place educators on a distribution of performance relative to other educators with students in the same grade and subject on the specific student assessment used as the outcome. The value of VAMs depends in significant part on the validity of the underlying student assessments in capturing what students ought to be learning. Because VAMs are not measures of student achievement growth, they cannot measure growth with respect to the Pennsylvania Academic Standards. VAMs measure the difference between actual and predicted scores for outcomes that are, at best, proximal measures of academic standards.

2. Advantages and Limitations

VAMs have been studied extensively and have been the subject of considerable policy discussion at the local, state, and national levels. The policy interest in value-added has risen recently in response to the U.S. Department of Education's Race to the Top (RTTT) initiative, which makes

competitive grants to states that agree to make student achievement part of annual evaluations of teacher and principal effectiveness. A recent issue brief found that eight states and the District of Columbia recently enacted new legislation to make student performance a major component of evaluations for general education teachers (Pennsylvania Clearinghouse for Education Research, 2011). Many states mandate that half of a teacher's evaluation must depend on student achievement (accounting for prior achievement).

To facilitate a broader understanding of value-added and its potential use as a component in teacher and principal evaluations, we list key strengths and limitations of the approach in Table I.1. In addition, in September 2010 Mathematica conducted a synthesis of information on the research and implementation of VAMs for Team PA (Lipscomb et al. 2010b).[6] In that review, we selected 21 studies that represent key issues and results in the literature and examined varying degrees of value-added implementation in seven school districts or states.

Table I.1. Strengths and Limitations of Value- Added Models Relative to Other Evaluation Methods

Strengths	Limitations
Focuses on outcomes rather than practice so it might encourage educators to better tailor practice to student needs	Restricted to effectiveness as measured through outcomes that can be systematically measured
Provides an objective measure of performance at the level of the individual teacher or school	Applied only in tested grades and subjects
Produces estimates of educators' contributions to achievement growth that account for students' starting points and other observed characteristics	Connection between school value-added and principal effectiveness is unclear
Results known to differentiate among staff at least at the tails of the performance distribution	Communicating statistical methodology to nontechnical audiences can be difficult

Sources: Pennsylvania Clearinghouse for Education Research (2011) and Mathematica.

The research synthesis highlighted several general findings that were used, in turn, to inform the goals and subsequent analyses undertaken for this report. We found consistent support for the existence of a wide distribution of teacher effectiveness with respect to student test score growth. As one might expect, teacher quality is the most important school-based factor affecting students. In most studies, the top 15 percent of math and reading teachers were capable of raising the achievement of the median-performing student at least 5 to 8 percentile points with one year of teaching compared with the teacher with the median value-added score.

We also found that few research studies examined the application of value-added to principals, although numerous studies examined its application to teachers and schools.[7] Due to the scarcity of research on principal value-added, we investigate in this report whether the average contribution to student achievement among educators at a principal's school approximates the principal's contribution, as the two should not be presumed to be synonymous. We ultimately conclude that,

[6] The review is available online through Mathematica's web site at [http://www.mathematica-mpr.com/publications/PDFs/education/teacherprin_valueadded.pdf].

[7] Dhuey and Smith (2011) is a recent addition to the principal value-added literature.

for many principals, it is impossible to distinguish the principal's contribution to student achievement from the contribution of other facets of the school (notably including the collective contribution of teachers). In consequence, throughout the report we label the principal-based measure as an estimate of the value-added of the *principal's school,* rather than an estimate of the value-added of the principal.

As indicated by Table I.1, value-added provides an objective measure of individual performance but one focused narrowly on test scores. The need to rely on assessment data has proven to be a practical challenge in extending value-added to an entire teaching staff. This limitation underscores the importance of determining through the Pennsylvania pilot study whether certain teacher or principal practices that can be measured through classroom observations in all grades and subjects are strongly tied to larger contributions to student achievement growth in tested grades and subjects.

The research literature also makes clear that an evaluation system can be considered fair only if it is based on valid and reliable measures. By validity, we mean whether the evaluation model measures what it intends to measure or whether it systematically over- (or under-) estimates performance for some teachers or principals. By reliability, we mean whether repeated measurements lead to a consistent result. Critics of value-added have voiced concerns that it is a noisy signal and that any of a litany of important factors can lead to the misclassification of some teachers as high or low performers (for example, nonrandom assignment of students into classrooms, small samples, incomplete statistical controls, or assessments that do not reflect the curriculum or standards). These concerns should not be swept under the rug. At the same time, we feel that they are not reasons to discard value-added analyses entirely. We share the view of a recent Brookings task force comprised of national experts on teacher quality in arguing that the best response is "to improve value-added measures continually and to use them wisely, not to discard or ignore the data" (Glazerman et al. 2010).

When the outcome is student test scores, value-added has been shown to be a better indicator of teacher effectiveness than teacher graduate degrees, certification, and experience after the initial five years of service (Goldhaber and Hansen 2010). Glazerman et al. (2010) also caution against setting unrealistic expectations for value-added as a performance measure, pointing out that the year-to-year correlation of value-added estimates for teachers—though modest—is as good as what has been found for measures used to make high-stakes decisions in other occupations. Value-added almost certainly provides better information for evaluating teacher and school effectiveness when compared against the alternative of maintaining the current system of evaluation in many school districts and states. In 2011, PDE found that 99 percent of teachers in the Commonwealth received a satisfactory rating for the 2009–2010 year (Team Pennsylvania Foundation 2011). In other words, the current system differentiates only a very small number of teachers with the absolute lowest ratings. Improving the evaluation framework will involve increasing the ability to differentiate high and low performance. It will also require ensuring that raters are trained to implement the new framework consistently for any new system to be deemed fair (Lane and Horner 2011).

In the following chapters, we present findings from analyses that address the study's three research questions. In Chapter II, we describe characteristics of the VAMs, such as the outcome measures, control variables, and applicability of estimates to Phase 1 teachers. We then present findings pertaining to teacher effectiveness measures using state-mandated and other assessments in Chapter III. In Chapter IV, we characterize relationships between teacher effectiveness and teacher practices to the extent possible in the Phase 1 pilot sample. We then present findings pertaining to principal and school effectiveness measures based on assessment and non-assessment data in

Chapter V. Finally, we provide a brief conclusion in Chapter VI with recommended next steps for this strand of the pilot study in subsequent phases. Interested readers are directed to Appendices A through C for technical information on the methodology, samples, and results, respectively.

II. CHARACTERISTICS OF VAMS ESTIMATED IN THIS REPORT AND THE APPLICABILITY OF EFFECTIVENESS MEASURES TO TEACHERS IN PHASE 1

The value-added models (VAMs) for this report include different outcome measures, control variables, and student samples. In this chapter, we provide a nontechnical description of the characteristics of the VAMs for teachers and schools that produce the results we discuss later in the report. We list the outcome measures, prior achievement controls, other background variables, and student samples that are included. We also show the extent to which Phase 1 teacher have at least one VAM estimate from across outcomes and therefore can be included in the analysis that examines relationships between value-added and observation-based measures.

A. Outcome Measures

We selected outcome measures for this report using the following two criteria that reflect goals for the pilot analysis:

1. The set of outcomes should include multiple measures of student outcomes, including non-PSSA test-based measures and nontest measures.
2. The value-added estimates based on the set of outcomes should include as many teachers from Phase 1 as possible with at least one estimate.

These selection criteria are consistent with the purpose of a pilot study in which findings are used to inform future development work and have no actual consequences for teachers, principals, or schools. In deciding whether to include specific outcomes, we did not assess the degree to which the measures correspond to the content that teachers are asked to teach or to which scores are indicators of skill acquisition by students. Our focus was in estimating VAMs to assess the extent to which attributions to teachers or principals are feasible. We withhold judgment on whether specific outcomes should or should not be included in Pennsylvania's model statewide evaluation system. Deciding which outcomes to include in the actual evaluation model will involve policy discussions that are outside the scope for Phase 1 (for example, discussions about a measure's degree of alignment with curriculum and standards, its validity and reliability, whether it is administered to all students or only to some students in a grade, the extent to which scores are malleable, and whether/how to allow for discretion at the district level in selecting measures).

In Table II.1, we list the student outcomes that are used in the primary VAM calculations for this report. The test-based outcomes come from the Pennsylvania System of School Assessment (PSSA), from Allentown's Progress Assessment (Progress), and from the Dynamic Indicators of Basic Early Literacy Skills (DIBELS). The PSSA is the statewide assessment that is given to all students in grades 3 to 8 and 11. It is also used for compliance with federal school accountability policies. Progress Assessments are curriculum-based measures that were developed by teacher committees in Allentown. They are administered to students multiple times during the year and are cumulative up to the date they are given. DIBELS includes several diagnostic measures that teachers can use to monitor students' early literacy and early reading skill development. The nontest outcomes include a student's rate of attendance and a measure that we constructed and refer to as

holding power.[8] The holding power variable is a binary measure of whether high school students during a given academic year enroll in any Pennsylvania public school the following year, which we interpret as an indicator of students who are likely to complete high school.[9] Although it will overestimate dropout rates across the state (because some students disappearing from the data are enrolled in private schools and others are enrolled outside the state), we expect it to permit a fair comparison among schools. We include attendance and holding power outcomes for school evaluation but not for teacher evaluation because these measures are likely to be affected by multiple staff at the building level.

Table II.1. Outcomes Considered in Value- Added Models for Teacher and School Evaluation in this Report

Outcome	Subject(s)	Grade	Teacher Evaluation	School Evaluation	Cohorts
PSSA (scaled score)	M, R	3	A, C, M	A, C, M	1
PSSA	M, R, S	4	PA	PA	3
PSSA	M, R, W	5	PA	PA	3
PSSA	M, R	6	PA	PA	3
PSSA	M, R	7	PA	PA	3
PSSA	M, R, W, S	8	PA	PA	3
PSSA	M, R, W, S	11	A, C, M	PA	1 (T); 2 (P)
Progress (raw score)	W	1	A	A	1
Progress	M, W	2	A	A	1
Progress	W	3	A	A	1
DIBELS (raw score)	R (NWF, PSF)	1	A	A	1
DIBELS	R (ORF)	2	A, C	A, C	1
Attendance (%)	--	4-12	--	A, M, N	1
Holding Power {0,1}	--	9	--	PA	3
Holding Power	--	10	--	PA	2
Holding Power	--	11	--	PA	1

Note: VAMs based on PSSA scores include students taking the modified version of the PSSA.

Subject abbreviations: M = Math; R = Reading; S = Science; W = Writing. DIBELS abbreviations: NWF = nonsense word fluency; ORF = oral reading fluency; PSF = phoneme segmentation fluency. Sample abbreviations: A = Allentown; C = Cornell; M = Mohawk; N = Northwest Tri-County; PA = Pennsylvania. T = Teacher; P = Principal.

-- indicates outcomes that are not specific to a particular academic subject or for teacher evaluation.

Table II.1 also indicates the subjects, grade levels, samples, and number of student cohorts for each VAM. All VAMs are estimated separately by subject and grade except for the attendance rate VAM. We included multiple grades of attendance information together to maximize sample sizes

[8] Some researchers use the term *value-added* only when there is a baseline measure of the outcome. We use the term for models without baseline measures of the outcomes because the methodology is very similar; in particular, it still involves comparing actual and predicted values of an outcome.

[9] In the VAMs for holding power, a student's enrollment decision in the following year is attributed only to the school that a student attends in the current year. This approach ignores any lingering effects of a student's previous schools. However, this approach is consistent with all other types of teacher and school VAMs that attribute a student's current-year test score growth only to the effects of the student's current-year teachers or schools.

within the Phase 1 districts, because attendance data were not available to us statewide. For Pennsylvania's evaluation system, our preference is for statewide samples whenever possible because the findings are the most inclusive; VAMs based on just a subset of districts are representative only of the districts they include. The lack of a statewide sample for the attendance VAM thus underscores an important point about VAMs: *The viability of any outcome measure in a VAM relies fundamentally on its availability across students who are relevant for the analysis.* When statewide samples were not available for other outcomes, we took the same approach of requesting the information directly from any Phase 1 district that collected it.

For three sets of PSSA models—grades 3 and 11 for teachers and grade 3 for schools—we are limited to pilot district samples even though the measures are collected statewide. Because grade 3 is the first year of state-mandated testing, there is no available baseline achievement measure that is collected across the Commonwealth. In order to include grade 3, we obtained student data on the fall administration of the grade 3 4Sight assessment from Phase 1 districts and used those scores to control for students' baseline achievement levels. A related problem affects the VAMs in grade 11 because students are not assessed statewide in grades 9 and 10. For the school VAMs, we are able to preserve the Pennsylvania sample by using students' grade 8 scores as their baselines, thus measuring contributions to achievement between grades 8 and 11. This is allowable for principal–school models because students are typically served by the same school during high school grades. Teachers, however, affect students in the year they educate them, making it critical to establish a baseline either at the end of grade 10 or at the beginning of grade 11. As we describe later in this chapter, we used fall 4Sight scores from Phase 1 districts, thus limiting the student sample to those districts.

In the final column of Table II.1, we show the number of student cohorts in each VAM. By cohort, we mean all the students a teacher educates or all the students attending a principal's school during an academic year. Incorporating multiple cohorts of students into a VAM can improve both the validity and the reliability of the estimates by averaging out random year-to-year fluctuations in student performance that affect teacher or school estimates from a single year of data (Schochet and Chiang 2010). Koedel and Betts (2011) showed that multiple cohorts improve validity as well because systematic biases offset one another over multiple years. Our primary models include all available student cohorts, up to three, moving backward in time from the most recent school year. For example, the three-cohort teacher VAM for grade 4 math includes all students a teacher taught in math between 2008–2009 and 2010–2011 who took the grade 4 math PSSA. For several outcomes, only one or two cohorts of students can be included using the data that we can access currently.[10] Estimates based on VAMs that include fewer student cohorts will be measured with greater noise, but they also have the advantage of better reflecting immediate past performance.

In the future, the Pennsylvania Department of Education (PDE) might wish to pursue a different set of outcome measures, including measures not included in this report. We focused narrowly on the academic subjects covered by the Phase 1 pilot (that is, math, English-language arts, and science). We also considered—but ultimately did not pursue—models based on the

[10] School VAMs based on grade 11 PSSA data include two cohorts because we use students' scores from three years earlier as their baseline scores and have data back only to 2006–2007. One cohort of student data is available for outcomes based on pilot samples. The number of cohorts in holding power VAMs differs by grade. In future years, three cohorts will be available for all grades, including grade 12.

Pennsylvania Alternate System of Assessment (PASA), the 4Sight, and core course passage rates. The PASA is given to students with severe cognitive disabilities instead of the PSSA if specified by their Individualized Education Program. On average, there are one or two PASA students per school and grade in Pennsylvania. Consistent with other studies, we report estimates for individual teachers and schools only when they are based on more than 10 students. Thus, the PASA data would not have been sufficient to estimate impacts for most teachers or schools in our sample. Moreover, at the school level, we found in exploratory work that including PASA only marginally increased the number of schools in Pennsylvania with at least one VAM estimate above the number obtained through the PSSA alone. Finally, there are technical issues related to involving the PASA that would be too resource-intensive to resolve for this report given the possible benefits of including it.[11] We thus exclude this measure and do not include students with severe cognitive disabilities in this report. However, we are able to include the vast majority of student with disabilities because most of them take either the PSSA or the modified version of the PSSA.

The 4Sight is a quarterly formative assessment that is intended for teachers as a low-stakes diagnostic indicator of student performance on content that mirrors the PSSA. We did not include it as an outcome measure (despite including it as a baseline measure for some VAMs) because it is given in the same subjects and grades as the PSSA, therefore meaning that it would not augment the coverage of value-added estimates to teachers. We prefer the PSSA as a measure because it is already used for school accountability, suggesting that teachers are motivated to have their students perform well on that test. Lastly, we examined the potential to use core course completion rates as a nontest outcome at the high school level. Though the data were available in the Phase 1 districts, we did not include those data because the small size of the Phase 1 pilot meant that we would not be able to present the findings without inadvertently identifying some schools.

B. Teachers with VAM Estimates from Phase 1 of the Pilot

Using the assessments listed in Table II.1, we were able to cover slightly more than half of the 153 teachers who participated in Phase 1 with at least one value-added estimate. Each teacher was observed by his or her principal in one grade and subject. Classroom observation data were not collected on pilot teachers in multiple subjects and grades even though the teachers might educate students in multiple subjects and grades. In Table II.2, we show how the pilot teacher sample was distributed across grades and subjects. The sample was selected by Dr. Suzanne Lane at the University of Pittsburgh, with input from Mathematica and superintendents in the pilot districts. It was limited to math, English-language arts, and science because assessment data are most often available in these subjects. Grades were selected to be representative of the K–12 spectrum. The sample sought to balance PSSA-tested grades and subjects and other grade/subject combinations in which the PSSA is not administered. More than half of the sample came from Allentown due to that district's size relative to the others.

[11] The VAM would have to account for how the PASA is reported in Pennsylvania's longitudinal student data on a categorical, rather than continuous, scale and is administered at three different levels of difficulty. Furthermore, there are substantial sample-selection concerns related to treating students who alternate between taking the PASA and a version of the PSSA in different years.

Table II.2. Summary of Teacher Evaluation Pilot, Phase 1

Subject	Grade 1	Grade 2	Grade 3	Grade 4	Grade 5	Middle School	High School	Total
Math	6	4	8	0	11	14	16	59
English-Language Arts	10	7	8	0	7	16	12	60
Science	0	5	0	12	0	8	9	34
Total	**16**	**16**	**16**	**12**	**18**	**38**	**37**	**153**

Note: Participants included teachers from the following school districts: Allentown (84), Mohawk Area (39), Cornell (20), and Northwest Tri-County IU5 (10).

The final sample of the 153 teachers for Phase 1 included 79 fewer teachers than had originally been selected across these same grades and subjects. The sample reduction was due primarily to the loss of one school district and one charter school that were slated to participate. We doubt that the sample loss affected our success rate in mapping VAM estimates to participating teachers because these 79 teachers had been assigned fairly evenly across grades and subjects.

The more serious concern for the pilot is that the Phase 1 sample is under-powered. Only 81 Phase 1 teachers have a VAM estimate that we can use in Chapter IV for studying relationships between teacher practices and larger individual contributions to student achievement. Based on this sample size, we can detect at best a 0.30 correlation between value-added and classroom observation scores.[12]

In Table II.3, we report the number and percentage of Phase 1 teachers with at least one VAM estimate from the analyses undertaken for this report. Overall, 53 percent of Phase 1 teachers have at least one VAM estimate that can be included in the Chapter IV analyses. When a value-added estimate could not be assigned, it was for one of two primary reasons. First, assessments were not always available in subject/grades/districts covered by the pilot (for example, second-grade science, or second-grade math outside of Allentown). Second, teachers did not always educate more than 10 students with an assessment score in the subject for which they were observed—a minimum number of students that we specified based on prior studies reporting estimates that are not overly noisy due to small sample sizes. This latter constraint affected all Phase 1 teachers in Northwest and many in Allentown who teach primarily special education students and students for whom English is a second language.

[12] This power calculation assumes a power level of 0.80 and a 5 percent confidence interval.

Table II.3. Number of Teachers from Phase 1 with at Least One VAM Estimate that Can Be Used for Correlating Value- Added with Teacher Practices in Chapter IV

Subject	Grade 1	Grade 2	Grade 3	Grade 4	Grade 5	Grade 8	Grade 11	Total
Math	0	1	6	0	9	5	7	28
English-Language Arts	5	5	7	0	6	5	5	33
Science	0	0	0	11	0	4	5	20
Total	**5**	**6**	**13**	**11**	**15**	**15**	**17**	**81**

Finally, we were not able to use the data on any teacher in Cornell because the evaluators in that district assigned all Phase 1 teachers exactly the same score on all rubric items, meaning that there is no variation in the classroom observation data for this district. In principle, these teachers could still be included in the sample for studying relationships between teacher practices and teacher effectiveness as measured by value-added because their data are complete. But preserving them in the sample does not contribute any information to the analysis; rather, it adds only noise. The teacher counts in Table II.3 reflect the deletion of the Cornell teacher sample. Omitting these teachers, we are able to use data on 61 percent of the remaining Phase 1 teacher sample in the Chapter IV analysis.

C. Control Variables that Are Included in the VAMs

1. Baseline Student Achievement

All VAMs used in education make predictions about student performance based on students' own achievement histories. Most researchers include prior scores from multiple academic subjects regardless of the subject of the outcome measure. We selected baseline measures—listed in Table II.4—by following a two-part strategy that is applied to each VAM based on the particular sample used:

1. Include scores from all available subjects in either the fall of the current grade or the spring of the prior grade—treating grade 8 scores as prior-grade scores for grade 11 students and showing preference for including measures that would allow for a statewide analysis if possible.
2. Include a same-subject PSSA score from two prior grades if one is available—substituting math for science scores and reading for writing scores

Table II.4. Baseline Measures for Value- Added Models Estimated in this Report, by Outcome

Outcome	Subject(s)	Grade	Sample	Prior Grade Baselines	Other Baseline Controls
PSSA	M, R	3	A,C,M	--	4Sight, fall Gr. 3 (M, R)
PSSA	M, R, S	4	PA	PSSA, Gr. 3 (M, R)	--
PSSA	M, R, W	5	PA	PSSA, Gr. 4 (M, R, S)	PSSA, Gr. 3 (M or R)
PSSA	M, R	6	PA	PSSA, Gr. 5 (M, R, W)	PSSA, Gr. 4 (M or R)
PSSA	M, R	7	PA	PSSA, Gr. 6 (M, R)	PSSA, Gr. 5 (M or R)
PSSA	M, R, W, S	8	PA	PSSA, Gr. 7 (M, R)	PSSA, Gr. 6 (M or R)
PSSA (Teacher)	M, R, W, S	11	A,C,M	--	4Sight, fall Gr. 11 (M, R) PSSA, Gr. 8 (M, R, W)
PSSA (School)	M, R, W, S	11	PA	--	PSSA, Gr. 8 (M, R, W)
Progress	W	1	A	--	Progress, fall Gr. 1 (W); DIBELS, fall Gr. 1 (NWF, PSF)
Progress	M, W	2	A	--	Progress, fall Gr. 2 (M, W); DIBELS, fall Gr. 2 (ORF)
Progress	W	3	A	--	Progress, fall Gr. 3 (W); 4sight, fall Gr. 3 (M, R)
DIBELS	R (NWF, PSF)	1	A	--	Progress, fall Gr. 1 (W); DIBELS, fall Gr. 1 (NWF, PSF)
DIBELS	R (ORF)	2	A,C	--	DIBELS, fall Gr. 2 (ORF)
Attendance (%)	--	4-12	A,M,N	Attendance, Gr. 3-11; PSSA, Gr. 3-11 (M, R)	--
Holding power {0,1}	--	9-11	PA	PSSA, Gr. 8 (M, R, W)	--

Notes: Baselines are given in the spring of the prior grade unless otherwise indicated.

Subject abbreviations: M = Math; R = Reading; S = Science; W = Writing. DIBELS abbreviations: ORF = oral reading fluency; NWF = nonsense word fluency; PSF = phoneme segmentation fluency. Sample abbreviations: A = Allentown; C = Cornell; M = Mohawk; N = Northwest Tri-County; PA = Pennsylvania.

Controls for prior achievement are the most important factors in any VAM because a student's own achievement history is the most important factor by far in predicting actual achievement at the end of the year—much more important statistically than the contributions of the teacher in any single year. Adding more extensive controls for prior student achievement can provide for better predictions about student achievement, enhancing the internal validity of effectiveness measures. But it typically comes at a cost of excluding students who lack scores on the additional assessments for which controls are being added. In other words, more extensive controls can yield better indicators of teacher effectiveness but the indicators are applicable to a smaller number of the teacher's students. This tradeoff between sample size and greater controls is clearest around the issue of whether to include a score from two prior grades ago (that is, grade 4 for 6th graders) because it has implications for whether mobile students can be included. We opted for the greater internal validity because we found that the direct sample loss was only 5 percent.[13] However, we note that the students who are dropped might not be a random sample of students, as they could differ relative to other students on characteristics beyond their mobility.

[13] The sample loss is lower than what would be found in the data systems of individual school districts because the statewide data retain the achievement histories of students who move between districts in Pennsylvania.

2. Additional Student- and Classroom-Level Variables

Along with controlling for baseline student achievement, most VAMs account for observable student background characteristics to help isolate further the contributions of educators to student achievement. The factors that are included in the VAMs are thought to be correlated with student performance while also being outside the control of teachers and schools. The standard list of controls would include measures related to students' socioeconomic status (for example, parent educational attainment, family income, or proxies such as eligibility for the free or reduced-price meals programs); family structure (for example, living in a single-parent household); or eligibility for programs such as special education. Unfortunately, there is usually a discrepancy between the variables that ideally would be included and the variables that are available in the data system. Researchers and policymakers are then left with a very difficult choice between estimating a model that could systematically over- or under-estimate teacher contributions due to less-than-complete controls and attempting to compensate at least partially for the omitted variables by including other measures that are available in the data. In practice, most data systems collect only limited information on student background characteristics, typically basic demographic variables such as gender, race/ethnicity, meals program eligibility, disability status, and English-language learner (ELL) status. Ultimately, most researchers and policymakers opt to include whatever information is available. At the same time, they acknowledge that a different set of variables would be preferable. The unavailability of student background controls is a difficulty in the short run, but data systems can be expanded over time to allow for a different set of variables to be used.

We adopted this same approach of including measures that are available in the data system both because we find that they are significant predictors of student performance and because there is a foundation for including them in prior research studies (Lipscomb et al. 2010b). The measures, listed in Table II.5, include variables for meals program eligibility, ELL status, categories of disability, mobility, grade repetition and age, flags for the modified version of the PSSA, gender, and race/ethnicity. We are not able to control for other measures of socioeconomic status, measures of family structure, or prior rates of student attendance in the data available.

Table II.5. Student and Classroom Control Variables Included in VAMs Estimated for this Report

Control Variable	Definition	Used in VAMs for Schools	Used in VAMs for Teachers
Free Meals	Free meals eligibility {0,1}	√	√
Reduced-Price Meals	Reduced-price meals eligibility {0,1}	√	√
English-Language Learner (ELL)	ELL in outcome year {0,1}	√	√
Specific Learning Disability (SLD)	Designation of SLD under IDEA {0,1}	√	√
Speech or Language Impairment (SLI)	Designation of SLI under IDEA {0,1}	√	√
Emotional Disturbance (ED)	Designation of ED under IDEA {0,1}	√	√
Intellectual Disability (ID)	Designation of ID under IDEA {0,1}	√	√
Autism (AUT)	Designation of AUT under IDEA {0,1}	√	√
Physical/Sensory Impairment	Designation of hearing impairment, visual impairment, deaf-blindness, or orthopedic impairment under IDEA {0,1}	√	√
Other Impairment	Designation of other health impairment, multiple disabilities, developmental delay, or traumatic brain injury under IDEA {0,1}	√	√
Mobility	Attended multiple schools during school year {0,1}	√	√
Grade Repeater	Repetition of the current grade {0,1}	√	√
Behind	More than 1.5 years older than expected for grade {0,1}	√	√
Age	Student age in years as of September 1	√	√
PSSA-Modified (outcome)	Outcome is a PSSA-M score (PSSA outcomes only) {0,1}	√	√
PSSA-Modified (baseline)	Baseline is a PSSA-M score (PSSA baselines only) {0,1}	√	√
Gender	Female {0,1}	√	√
Race/Ethnicity	Indicators for African American, Hispanic, Asian Pacific Islander, or other race/ethnicity {0,1}	√	√
Classroom-Level Characteristics	Separate classroom-level variables for free meals, reduced-price meals, ELL, special education, gender, and race/ethnicity (percentage of students in the classroom)		√
Classroom Size	Number of students in the classroom		√
Classroom Size Interactions with Student-Level Characteristics	Separate interaction terms between classroom size and the following student-level characteristics: ED, ID, AUT, physical/sensory impairment, free meals, and ELL		√

Note: The value of a classroom-level variable for a particular student is an average across the courses that a student takes in the subject assessed by the outcome measure during the year.

Abbreviation: IDEA=Individuals with Disabilities Education Act.

Among these variables, the inclusion of gender and race/ethnicity controls is the most controversial. The intent is not to set different standards for students. Rather, it is an empirical acknowledgement that in the absence of preferable measures, these variables explain a statistically significant portion of the variation in student performance even after accounting for prior student achievement and all the other variables in Table II.5. To the extent that gender and race/ethnicity

represent unobserved factors that differ across students and are outside the control of teachers and schools, the VAM estimates would systematically penalize or advantage certain teachers and schools if these controls were omitted. If fuller controls were available, we would expect that the amount of variation that gender and race/ethnicity control for would shrink and eventually become statistically insignificant.

In addition to the student-level variables that we include in all teacher and school VAMs, we also include several classroom-level variables in teacher VAMs that account for peer influences on achievement.[14] These measures are intended to account for various inputs that are largely beyond the control of teachers but affect their overall workload. The controls include the average characteristics of students in the classroom, the classroom size, and interaction variables between classroom size and student characteristics that indicate more severe needs. When a student takes multiple courses during the year in a subject, the peer variables are averaged across classrooms. We do not include classroom-level variables in the school VAMs because the make-up of classrooms is within the control of school administrators. We also do not include any measures related to educators' own characteristics (for example, their years of experience) that might affect their effectiveness relative to other educators.

We adjust some teacher effectiveness measures by subtracting the average value-added score among teachers in a school or district. This type of adjustment has the potential to control better for school- or district-level influences that affect the performance of all teachers at a school or district. But it also has two implications that might be disadvantageous for a statewide evaluation system. First, this type of adjustment changes VAM inferences so that teachers are compared only with other teachers in their same district or school, rather than with other teachers in the state. Second, it might under-represent true differences in teacher contributions across districts and schools if highly effective (or ineffective) teachers tend to cluster. For these reasons, we use the adjusted estimates for assessing the extent to which the variation in teacher effectiveness is primarily across or within districts and schools but not as a part of our primary VAMs. We believe that identifying district- or school-level variables that could control for this variation without preventing a statewide comparison of effectiveness would be very informative in follow-up work.

[14] The exception is for VAMs that include a single cohort of students in elementary grades. Because classrooms in elementary grades tend to be self-contained, it is not possible to separate a teacher's influence from the influence of students' peers with one year of teaching data.

III. VALUE- ADDED RESULTS FOR TEACHERS

In this chapter, we present findings from value-added models (VAMs) that are intended to produce measures of teacher effectiveness. Our focus is on characterizing the distributions of teacher effectiveness across outcome measures, subjects, and grades. We begin by discussing teacher quality distributions in selected grades and subjects based on Pennsylvania System of School Assessment (PSSA) scores, and we conclude that sizable differences in quality exist across Pennsylvania. We then describe issues related to statistical uncertainty and how VAMs quantify the extent of imprecision through confidence intervals. Next, we contrast the teacher effectiveness estimates with estimates obtained through several alternative specifications to examine the sensitivity of results. The alternative specifications adjust the VAMs for district and school factors, omit a same-subject prior score control, and control for prior achievement using a beginning-of-year score rather than an end-of-year score, respectively. In the final section of this chapter, we describe several key characteristics of the teacher effectiveness estimates generated by estimating VAMs on additional outcomes and student samples from the pilot districts.

A. Variation in Teacher Effectiveness Based on PSSA Outcomes

Consistent with findings on teacher quality in the research literature, we find sizable differences in teacher effectiveness across Pennsylvania, as measured by value-added in math, reading, and science. In Table III.1, we depict the variation in teacher effectiveness based on PSSA scores for three subject-grade combinations covered by the Phase 1 pilot (that is, 5th-grade math, 8th-grade reading, and 4th-grade science). Teacher impacts are reported in terms of PSSA scaled scores at different points in the teacher quality distributions. The table values represent the expected difference in scaled scores between students educated by a given teacher and students educated by the average-performing Pennsylvania teacher, controlling for the factors described in Chapter II.

Table III.1. Distribution of Teacher VAM Estimates for Selected PSSA Outcomes

	Effectiveness of the Teacher at the Indicated Percentile Relative to the Effectiveness of the Average Teacher (in PSSA scale points)					
Outcome	5th	15th	25th	75th	85th	95th
Math PSSA, Grade 5	-70	-43	-28	+25	+44	+77
Reading PSSA, Grade 8	-35	-22	-15	+14	+22	+38
Science PSSA, Grade 4	-60	-39	-26	+24	+38	+68

Source: Mathematica calculations reported in Appendix Table C.1 based on Pennsylvania student data.

Note: Findings are based on a three-cohort model with statewide samples of teachers and students. The sample of teachers consists of those who served as teachers in every year from 2008-2009 to 2010-2011 in the outcome subject and grade, and their VAM estimates are based on students in their classrooms during that period.

PSSA = Pennsylvania System of School Assessment; VAM = value-added model.

In a single year of instruction by a teacher at the 15th percentile, a 5th-grade student originally at the median of the statewide distribution of math scores would score 43 scale points lower on the math PSSA than he or she would score with a single year of instruction by an average-performing teacher. On the other hand, this student would score 44 scale points higher by having the 85th percentile teacher than by having the average teacher. Thus, the 85th and 15th percentile teachers differ in their effectiveness by 87 PSSA points. This scale point difference can also be interpreted with reference to the statewide distribution of student test scores. By switching from the 15th

percentile teacher to the 85th percentile teacher, a 5th-grade student originally at the median of the statewide distribution of math scores would be predicted to rise to the 65th percentile.[15] Sizable variation in the teacher effectiveness estimates is also observed for other grade–subject combinations. The 85th and 15th percentile teachers differ in their estimated effects on PSSA scores by 44 points in 8th-grade reading and 77 points in 4th-grade science. The impact of one year of teaching by a teacher at the 85th percentile relative to a teacher at the 15th percentile is about 54 percent as large as the 2011 test score gap between African American and white students in 5th-grade math, 24 percent as large in 8th-grade reading, and 45 percent as large in 4th-grade science.[16]

Pennsylvania's most-effective teachers are certainly capable of moving the academic needle in the Commonwealth. However, teacher effectiveness is just one of many school-based and nonschool factors that affect students during the year. Students' own prior achievement scores are by far the most important predictors of their actual achievement scores. In Appendix Table C.2, we show estimated coefficients for the control variables that are included in the models reported in Table III.1. Baseline scores clearly have the most explanatory power, a finding that is common to all VAMs, not just to these three selected ones. The relationships between achievement and the other student- and classroom-level variables are typically statistically significant as well, due partly to a very large number of student observations, though they have much smaller magnitudes.

The teacher quality estimates that we find in Pennsylvania for math and reading are similar in size to those found by other researchers in different states and school districts (see Table III.2). We can compare the size of different teacher quality distributions by expressing the effectiveness estimates in terms of standard deviations of student scores relative to the average score. A standard deviation is approximately the amount by which the 85th percentile student score exceeds the 50th percentile score (or equivalently, it is approximately the amount by which the 50th percentile score exceeds the 15th percentile score). In Table III.2, we report teacher estimates from prior research in these terms for the teacher at the 85th percentile of effectiveness relative to the teacher at the 50th percentile. For example, the value of 0.20 would indicate that by switching from the 50th percentile teacher to the 85th percentile teacher, the score of a student originally at the median of the statewide distribution of scores would be predicted to rise by 0.20 standard deviations. This gain translates into an increase from the 50th percentile of student scores to the 58th percentile of scores. Smaller values in Table III.2 indicate that teachers are grouped tightly together in terms of their performance. Larger values indicate that teachers are spread farther apart.

[15] To make this calculation, we divided 87 scale points by 223 scale points, the standard deviation of 5th-grade PSSA math scores (see Appendix Table C.1). Thus, an 87 scale-point difference amounts to a difference of 0.39 standard deviations in the distribution of student scores. In the assumed normal distribution for student scores, moving from the 50th to the 65th percentile is equivalent to an increment of 0.39 standard deviations.

[16] The PSSA achievement gap between African American and white students in 2011 was approximately 160 scaled score points in 5th-grade math, 180 scaled score points in 8th-grade reading, and 170 scaled score points in 4th-grade science, among students with prior-grade scores in math and reading.

Table III.2. Teacher VAM Estimates in Recent Studies for the 85th Percentile of Effectiveness Relative to the 50th Percentile, Reported in Standard Deviations of Student Test Scores

Research Study	Math	Reading	Grade Range	Location
This study	0.16-0.23	0.09-0.16	4-8	Pennsylvania
Aaronson et al. (2007)	0.15	--	9	Chicago
Goldhaber and Hansen (2010)	0.22	0.10	4-5	North Carolina
Hanushek and Rivkin (2008)	0.13-0.20	--	4-8	Texas (1 large urban district)
Jacob and Lefgren (2008)	0.26	0.12	2-6	Western United States (1 midsize district)
Kane et al. (2008)	0.17-0.21	0.17-0.20	4-8	New York City
Kane and Staiger (2008)	0.16-0.19	0.13-0.16	2-5	Los Angeles, New York City, Boston
Koedel and Betts (2011)	0.18-0.24	--	4	San Diego
Lipscomb et al. (2010a)	0.15-0.20	0.11-0.14	4-8	Pittsburgh
Rothstein (2010)	0.15	0.11	5	North Carolina

Sources: Appendix Table C.1 and the individual articles, most of which are summarized in Lipscomb et al. (2010b).

Note: Findings from Lipscomb et al. (2010a) are for three-cohort VAMs using PSSA score outcomes.

PSSA = Pennsylvania System of School Assessment; VAM = value-added model.

We find that Pennsylvania teachers are capable of affecting test scores more in math than in reading, as indicated by larger standard deviations in math—a finding supported by external research studies.[17] We also find that distributions of effectiveness are larger in elementary grades than in middle school grades (Appendix Table C.1). Kane et al. (2008) found a similar pattern in estimating value-added in New York City, as did Lipscomb et al. (2010a) in Pittsburgh. The remaining studies in the table did not examine value-added distributions by grade level, but larger effectiveness distributions tended to be found in elementary grades across these studies.[18]

Teachers are capable of producing larger achievement gains in 4th-grade math than in 8th-grade reading (for example) partly because students tend to make relatively larger improvements in math and in elementary grades. A useful way to compare the size of teacher effects across subjects and grades is to adjust them for the average annual gains that we expect for students in each subject and grade. Then the estimates can be interpreted as an effect size relative to the amount of learning that we expect for typical students. We made this adjustment using the expected gains measures that are reported in Hill et al. (2008).[19] When averaged across grades and subjects, the results suggest that a typical student with a teacher at the 85th percentile learns roughly 40 percent more than other

[17] Similar distributions for science and writing in Pennsylvania are more like distributions for math than for reading. External estimates for science and writing are available only in Lipscomb et al. (2010a) for Pittsburgh. The results in that study are comparable to results in the present study.

[18] High school grades have been studied to a lesser extent, in part because state assessments are not usually administered in consecutive grades. Lipscomb et al. (2010a) present exploratory findings from VAMs based on 11th-grade PSSA outcomes and teaching data from 2009–2010. Relative to effectiveness distributions in grades 6–8 based on PSSA outcomes from 2009–2010, the 11th-grade effectiveness distributions were estimated to be similarly sized in math and smaller in reading.

[19] The adjustment measures are based on seven nationally normed tests. The denominators used for the adjustments are larger in math than in reading, and larger in earlier grades than in later grades.

students taught by the 50th percentile teacher in the same grade and subject during the year in terms of typical annual growth (see Appendix Table C.3). The impacts are similar across subjects in grades 4 and 5 but in grades 6–8 the impacts still tend to be larger in math than in reading.

B. Statistical Uncertainty in the Teacher Effectiveness Estimates

1. The Use of Confidence Intervals in VAMs

All performance measures are somewhat imprecise because they are based on limited information. To help quantify the precision of value-added measures, it is customary to report them together with a range of values called a confidence interval. In a hypothetical example, a teacher may place at the 45th percentile in terms of value-added in math with a confidence interval that ranges from the 40th to the 52nd percentile. Because the confidence interval includes the 50th percentile, the estimate of that teacher's effectiveness is no different statistically from average.[20] In reporting effectiveness, educators are treated as performing at the average unless their confidence intervals are entirely above or below the 50th percentile.[21]

The inclusion of a confidence interval is a reminder that value-added measures—like all other performance measures—are *estimates* of performance. There is debate about the size of the confidence interval that is acceptable for use in decision making, but the very reporting of confidence intervals is a distinct advantage of value-added measures over other measures for which a confidence interval is not reported. In classroom observation data, for instance, there is rarely an attempt to quantify the degree of imprecision around scores, although such an error band certainly exists. That is, if observations could be conducted many times for the same teacher in the same school year, the outcomes would probably differ based on factors such as the degree of reliability between different observers or even inclement weather that makes it difficult for students to concentrate on some days. Typically, however, only one evaluation score is obtained out of this distribution of possible scores. That score might over- or under-represent a teacher's typical performance. Value-added measures are confronted with related issues, but they can provide an indication about the degree to which a teacher's actual performance might be higher or lower than what it is estimated to be.

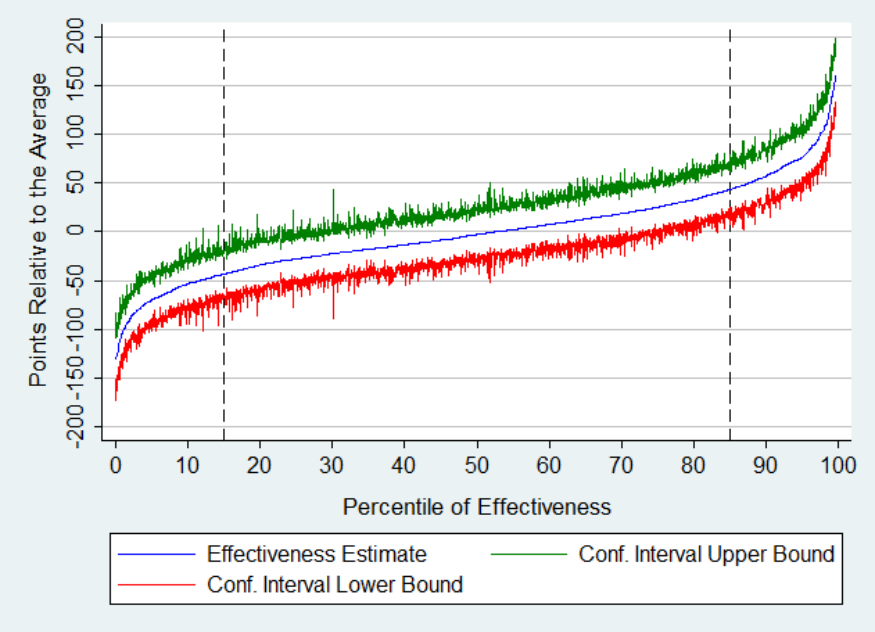
In Figures III.1 and III.2, we illustrate how confidence intervals are applied to value-added distributions in 5th-grade math and in 8th-grade reading, based on a teacher's students between 2008–2009 and 2010–2011. We do not illustrate the distribution of effectiveness scores in 4th-grade science because it is similar. In each figure, the horizontal axis indicates percentiles of teacher effectiveness and the vertical axis indicates the additional contribution of a given teacher relative to the average teacher in PSSA scale points. The blue curve depicts the effectiveness distribution based on all Pennsylvania teachers teaching in the grade and subject. The green and red scatters above and below the blue curve represent the bounds of the confidence interval for each individual teacher estimate. Teachers with confidence intervals that include a score of zero (the score achieved by the

[20] The 50th percentile is the median value. We refer to this value as the average because we expect the median and the average teacher effectiveness estimates to be very close if not identical.

[21] The statistical uncertainty of estimates relates directly to how much error there would be in classifying teachers or principals into performance categories on the basis of these estimates (Schochet and Chiang 2010).

50th percentile teacher) cannot be distinguished statistically from average. Overall, the performance of 52 percent of teachers in 5th-grade math and 30 percent of teachers in 8th-grade reading are statistically different from average based on students taught over the three-year period. (Appendix Table C.1). These tend to be the teachers who place at either end of the distribution. More teachers can be rated as above or below average in math because that subject has a larger distribution of estimates given the same-width confidence interval. Achieving an equal rate in reading would require even greater precision.[22]

Figure III.1. Distribution of Teacher Effectiveness and 95 Percent Confidence Intervals of Teacher Effectiveness Estimates for 5th- Grade Math PSSA Scores



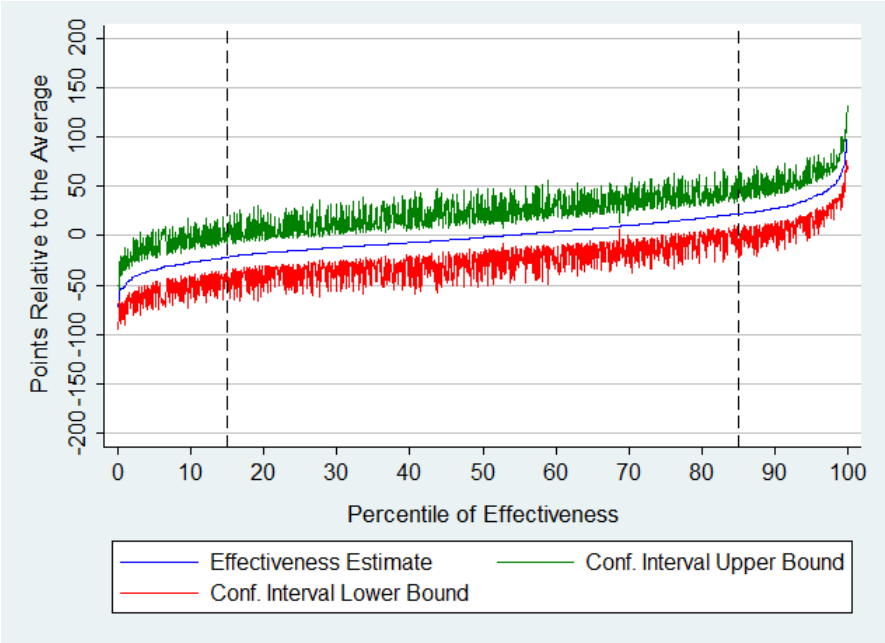
Source: Mathematica calculations based on data from PDE.

Note: Findings are based on a three-cohort model with statewide samples of teachers and students. The sample of teachers consists of those who served as teachers in every year from 2008-2009 to 2010-2011 in the outcome subject and grade.

PDE = Pennsylvania Department of Education; PSSA = Pennsylvania System of School Assessment.

[22] As we discuss next, adding more years of data would likely reduce the confidence intervals. However, this change might also reduce the amount of variation in the teacher effectiveness measures so the impact on the fraction of teachers that are statistically different from zero would be ambiguous a priori.

Figure III.2. Distribution of Teacher Effectiveness Estimates and 95 Percent Confidence Intervals of Teacher Effectiveness Estimates for 8th- Grade Reading PSSA Scores



Source: Mathematica calculations based on data from PDE.

Note: Findings are based on a 3-cohort model with statewide samples of teachers and students. The sample of teachers consists of those who served as teachers in every year from 2008-09 to 2010-11 in the outcome subject and grade.

PDE = Pennsylvania Department of Education; PSSA = Pennsylvania System of School Assessment.

2. Incorporating Students from Multiple Cohorts

Value-added models vary in terms of the number of student cohorts that they include, but it is common to include multiple cohorts whenever a VAM is used for a high-stakes purpose such as a performance evaluation. By incorporating data from multiple cohorts, we can reduce the size of confidence intervals (that is, improve precision) through using data on more students. Teacher effectiveness measures based on multiple student cohorts are averages of a teacher's contributions to students taught during the years that are considered. In any single year, a teacher's students can perform unexpectedly well or poorly on an assessment for reasons other than the teacher's direct contribution. Such random fluctuations would affect the teacher's effectiveness rating for that year. To the extent that these random fluctuations tend to average out over time, the multicohort VAM provides a more reliable measure of performance. Averaging effectiveness measures across multiple cohorts also can be advantageous for reducing the effects of systematic fluctuations in scores. As mentioned earlier, Koedell and Betts (2011) found that a three-cohort VAM can reduce the potential bias in teacher effectiveness estimates that is due to nonrandom assignments of students in teachers' classrooms to statistical insignificance. Finally, multicohort VAMs can also better distinguish teacher effects from the effects of students' peers in the classroom, which cannot be separately identified in a single-cohort model unless teachers teach in multiple classrooms during the year. For all of these reasons, our primary VAMs—including the three presented in the prior section—incorporate the three most recent student cohorts or up to three if fewer are available.

The decision to incorporate data from multiple student cohorts comes with tradeoffs in terms of not reflecting immediate past performance and yielding fewer teachers with estimates based on the full number of cohorts. By definition, a three-cohort VAM evaluates performance over a longer period than a one-cohort VAM. A three-cohort VAM will thus apply to fewer teachers if policymakers decide not to report estimates for teachers when they have data from only one or two prior cohorts. In Table III.3, we illustrate this tradeoff for the selected outcomes discussed here. The main columns show the number of teachers with estimates based on the full number of cohorts in each specification (that is, one or three), and the percentage of those estimates that are statistically different from average.

Table III.3. Number of Teachers with Effectiveness Estimates Reported and Share of Reported Estimates that Are Statistically Different from the Average, by Number of Cohorts Used in Estimation

	Number of Teachers with Estimates Reported		Percentage of Reported Estimates that Are Statistically Distinguishable from Average	
Outcome	1-Cohort Model	3-Cohort Model	1-Cohort Model	3-Cohort Model
Math PSSA, Grade 5	4,103	2,836	36.5	52.0
Reading PSSA, Grade 8	1,916	1,717	22.3	30.5
Science PSSA, Grade 4	4,187	2,854	27.7	49.8

Source: Mathematica calculations reported in Appendix Tables B.4 and C.1 based on data from PDE.

Note: Findings are based on statewide samples of teachers and students and a 95 percent confidence interval. The one-cohort model includes teachers with students in the outcome subject and grade in 2010-2011. The three-cohort model includes teachers with students in every year from 2008-2009 to 2010-2011 in the outcome subject and grade.

PDE = Pennsylvania Department of Education; PSSA = Pennsylvania System of School Assessment.

Moving from a one-cohort VAM to a three-cohort VAM decreases the number of teachers with effectiveness measures that are based on the full period that is considered but improves the precision of those estimates that are reported. For instance, the percentage of teachers with estimates reported in 5th-grade math declines by 30 percent but the share of them that are statistically significant increases by 40 percent. The gain in precision is not an artifact of using a different sample of teachers across specifications, because the percentages of statistically significant one-cohort VAM estimates among teachers who also have a three-cohort VAM estimate are nearly identical to the values reported in the table.

Adding additional cohorts of student data leads to a relatively larger reduction in the number of multicohort estimates based on the full panel of student cohorts for teachers in elementary grades than in middle school grades. This could at least partly be an implication of a requirement that we imposed whereby teachers have to be teaching students who take a particular subject and grade level assessment (for example, 5th-grade math PSSA) in *each* of the three prior school years. That is, elementary teachers who changed grade levels within the past three years would be excluded.[23]

[23] We would not actually have to limit the sample this way if the goal of the analysis was to calculate an overall value-added estimate for a teacher in a subject. In that case, we could require that a teacher have data from three prior years across grades and then calculate a composite estimate for that teacher.

Fewer middle school teachers would be affected by changes to their teaching assignments to the extent that such changes affect the number of classes in which a teacher instructs students in the subject and grade more than whether the teacher instructs any students at all in the subject and grade.

A compromise strategy (not considered for this report) could be to use three cohorts of student data in the VAM for a particular subject and grade level, and then report all estimates that apply to teachers with students in that subject and grade during the most recent year, regardless of whether they have been teaching that subject and grade for one, two, or three prior years. Under this approach, more teachers would have VAM estimates but the individual estimates would vary in terms of the number of student cohorts they include. This could be an attractive option if Pennsylvania wants to use value-added for teacher evaluations only when three years of data are available but also wants to provide value-added information to all teachers for diagnostic or professional development purposes. Decisions about the number of student cohorts to include in a VAM should be based foremost on the intended purpose of the VAM.

C. Sensitivity of Measured Effectiveness to Alternative VAM Specifications

The teacher effectiveness measures presented in this report depend on several model design elements in addition to choices about which outcomes, baselines, and student cohorts to include. We constructed the VAMs based on models that appear in the research literature but their features should be examined closely to ensure that they align with Pennsylvania's policy preferences. In this section, we explore the sensitivity of teacher effectiveness estimates to three alternative specifications of the VAM to illustrate the types of decisions that policymakers must consider in constructing an effectiveness measure. First, we adjust estimates for factors that might vary at the district or school levels. Second, we assess the likely performance of VAMs when a same-subject baseline score is not available, as in science. Third, we compare the impacts on teacher effectiveness estimates of controlling for prior achievement using a beginning-year score versus an end-of-year score from the prior grade. We find that effectiveness estimates from the primary model might not be highly sensitive to alternative specifications for most teachers although alternative specifications do affect the effectiveness estimate for some teachers.

1. Adjusting Measured Effectiveness for District or School Factors

The distinguishing feature of a VAM is its emphasis on separately identifying the individual contributions of educators to the achievement growth of their students. Some analysts have included school-specific indicators (that is, dummy variables)—in addition to variables measured at the student and/or classroom levels—to control for factors such as working conditions at the school that might affect both student performance and a teacher's ability to be effective in the classroom. When school-level indicators are excluded from VAMs (or similarly, when district-level indicators are excluded), teacher effectiveness measures incorporate any effect that schools have on student growth. This means that teachers at good schools (that is, those that improve student achievement more than others because of factors beyond the control of the teachers) will have an advantage in the sense that their estimated effects will be higher than similar teachers who teach at lower-quality schools. Adding school or district indicators factors out these across-school or across-district differences.

In a statewide evaluation system, however, including these indicator variables might not be desirable because the effectiveness estimates then implicitly compare teachers directly with other

teachers in the same school or in the same district rather than with other teachers in Pennsylvania. An above-average teacher who does not perform quite as well as his or her colleagues at a very high-performing school could actually be estimated to be below average in a model with school indicators. This would be especially undesirable if part of the reason a school performs well is the positive effect generated by having many good teachers. On the other hand, this could incentivize good teachers to move to bad schools and thereby promote equity. Factoring out average teacher impacts in a school can therefore lead to underestimating the teacher's influence if the average effect is simply due to the clustering of good (or bad) teachers rather than to a distinct school influence.[24] It also could undermine efforts to promote teamwork within a school because teacher effects would be measured only relative to other teachers in the same school or district. For these reasons, we do not include district or school indicators in our primary VAMs.

With controls for prior scores, student-level background characteristics, classroom-level characteristics, and teachers already added, it is possible that districts and schools do not have a substantial additional impact on student achievement. We can get an idea of the magnitude of the impact of omitting district and school effects in teacher VAMs by examining the change in standard deviation of estimated teacher effects after subtracting the within-district and within-school average effects from each estimate. We performed this analysis for math and reading by combining the teacher estimates from the three-cohort VAMs in grades 4 through 8, and then by subtracting the within-district or within-school average teacher effect from each individual teacher effect. Because the individual effectiveness distributions had different standard deviations, we first standardized them to a value of one before combining teachers across grades 4 through 8.[25] The adjusted distributions, summarized in Table III.4, consist of estimates that compare teachers with the average teacher in their district or school.

Table III.4. Implied Percentage of Variation in Teacher Value- Added Within Districts and Schools

Outcome	85th Minus 50th Percentile of VAM Estimates (in z-score units)			Implied Percentage of Total Variation in Teacher Value Added that Is Within Districts	Implied Percentage of Total Variation in Teacher Value Added that Is Within Schools
	Primary VAMs	Adjusted for Districts	Adjusted for Schools		
Math PSSA, Grades 4-8	1.00	0.91	0.77	83	59
Reading PSSA, Grades 4-8	1.00	0.93	0.81	87	65

Source: Mathematica calculations based on Pennsylvania data.

Note: Findings are based on three-cohort teacher VAM estimates for grades 4-8 that are reported in Appendix Table C.1. The implied percentage columns are calculated as the ratio of the square of each "adjusted" column value to the square of the corresponding primary VAM value. A z-score unit is a standard deviation of student scores.

PDE = Pennsylvania Department of Education; PSSA = Pennsylvania System of School Assessment.

[24] The contribution of principals is another school factor that might be confounded with effectiveness estimates for teachers. If teachers with high VAM scores simply serve at schools led by effective principals, the teacher's contribution might be less than what is measured by a model that excludes school indicators. However, because effective principals might recruit effective teachers, it is not clear that this variation should be removed from teacher effectiveness estimates.

[25] For teachers with students in multiple grades, their standardized grade-specific estimates were then averaged so that all teachers in grades 4 through 8 had one estimate for each subject. The district or school adjustment factor for an individual teacher is the average value-added across all teachers in the districts or schools where he or she teaches.

Most of the overall variation in Pennsylvania teacher effectiveness estimates in these grades and subjects is within individual schools rather than across them—a finding shared by many studies in the research literature. Most of the remaining variation is across schools within individual districts. The smallest portion of variation in Pennsylvania teacher effects is across districts. Specifically, we found that about 62 percent of the variation in the teacher effectiveness estimates is within schools, 23 percent is across schools within districts, and 15 percent is across districts. This is indicated in Table III.4 by averaging across rows the degree of remaining teacher quality variation in Pennsylvania after removing the average value-added of teachers in each district or school. For example, adjusting estimates from the primary models for districts leaves about 85 percent of the variation intact (that is, 83 percent in math and 87 percent in reading). Adjusting estimates from the primary models for schools reduces the amount of variation in the teacher effectiveness measures relatively more, but still 59 percent remains in math and 65 percent remains in reading.

These findings support a conclusion that the most important factors to include in a VAM are those that vary within schools. But the findings also indicate that about 38 percent of the variation in teacher effectiveness across Pennsylvania is across schools rather than within them. Thus, we cannot rule out the possibility that adding controls for certain school- or district-level factors could improve the validity of the estimates. An alternative method (not considered in this report) to adjusting for the average value-added in districts or schools that still accounts for the impact that districts and schools have on student achievement would be to include district- or school-level observable characteristics in the VAM. Examples of such characteristics could be the fraction of students eligible for free meals, the average years of experience among teachers, or the level of district funding per student. We did not pursue this approach for this report because we were concerned that the VAMs might not produce valid estimates for the relationships between school or district characteristics and outcomes. To estimate these relationships, the VAMs would have to rely only on year-to-year variation in the characteristics of the same school or district, which is much smaller and more transitory than the variation of interest across different schools and districts. Indeed, in preliminary analyses, we found that several coefficient estimates on student demographic variables measured at the school level had counterintuitive signs, which suggests that the VAMs might not produce valid estimates of these coefficients. We recommend that the Pennsylvania Department of Education (PDE) consider whether to control for factors that vary across districts or schools during Phase 2 of the pilot study and, if so, how best to do so.

2. Excluding a Prior Achievement Score from the Same Subject

Given the importance of students' own achievement histories in predicting current achievement, analysts seek to control for a prior assessment in which a student has a score in the same subject. Including a same-subject prior score is desirable but it is not a requirement for a VAM to operate because VAMs simply make a prediction about students' scores based on the factors that are controlled, whether they come from the same or different subjects. Because students are not always assessed in consecutive grades across subjects, a practical implication of extending the use of value-added broadly to teachers is that the model in some grades and subjects will not be able to incorporate a same-subject prior score.

An example comes from 4th-grade science. The VAM for 4th-grade science can only control for the incoming science abilities of students to the extent that they are related to prior achievement scores in math and reading, the two subjects tested by the PSSA in 3rd grade. Short of introducing a new 3rd-grade science assessment statewide, policymakers and analysts are left to decide whether to use the 4th-grade science VAM with available data or disregard it altogether. In the following

analysis, we infer the likely performance of the 4th-grade science VAM by simulating the impact on teacher effectiveness measures of intentionally omitting a same-subject prior score from VAMs for 5th-grade math and reading. We chose 5th grade for this diagnostic because 5th graders have three available scores from 4th grade: math, reading, and science. Specifically, we compared the math and reading teacher estimates obtained through VAMs that control for 4th-grade math and reading scores with estimates obtained by replacing the same-subject score with the 4th-grade science score. The rationale is that the former specification is what analysts and policymakers would like to estimate but the latter is equivalent to what can be estimated for 4th-grade science using available data.[26]

We show results from the exercise in Table III.5, which aggregates estimates for math and reading for presentation purposes. The table rows indicate teacher effectiveness quartiles from the specification that controls for math and reading scores in 4th grade. The columns indicate effectiveness quartiles from the specification that replaces the same-subject prior score (that is, 4th-grade math or reading, depending on the outcome) with the 4th-grade science score. The values indicate the number and percentage of teacher estimates in each cell. Teachers are included in the analysis only if they have a VAM estimate under both specifications, but most included teachers are represented twice because 5th-grade teachers typically teach students in both subjects.

Table III.5. Counts and Percentages of Grade 5 Math and Reading Teachers in Effectiveness Quartiles Based on 3- Cohort Teacher VAMs that Include and Exclude Same- Subject Baseline Scores

	Quartile of Effectiveness Based on Teacher VAM that Controls for the Grade 4 Science Score Instead of the Same- Subject Baseline Score				
	1st (bottom)	2nd	3rd	4th (top)	Total
Quartile of Effectiveness Based on Teacher VAMs with Controls for Grade 4 Math and Reading Scores					
1st (bottom)	1,085 (74.8)	310 (21.4)	54 (3.7)	1 (0.1)	1,450 (100.0)
2nd	314 (21.7)	727 (50.2)	375 (25.9)	32 (2.2)	1,448 (100.0)
3rd	49 (3.4)	363 (25.1)	758 (52.3)	279 (19.3)	1,449 (100.0)
4th (top)	2 (0.1)	48 (3.3)	262 (18.1)	1,136 (78.5)	1,448 (100.0)
Total	1,450 (25.0)	1,448 (25.0)	1,449 (25.0)	1,448 (25.0)	5,795 (100.0)

Source: Mathematica calculations based on Pennsylvania data.

Note: In each table cell, the first value is the number of teachers in the given cell, and the second value (in parentheses) is the percentage of the row total that is represented by that cell.

The teacher estimates show a relatively high degree of correlation across specifications, with most estimates falling on the table's diagonal elements. For instance, the top-left cell indicates that

[26] For the diagnostic purpose of these analyses, we did not include controls for 3d-grade scores.

75 percent of the teachers with the highest 25 percent of effectiveness scores under the primary model that includes prior math and reading scores had effectiveness scores in the top 25 percent under the alternative model too. Of the teachers whose quartile position changes, nearly all of them move by just one quartile. Only three of nearly 5,800 teacher estimates, or 0.05 percent, move from the top quartile to the bottom quartile. The within-teacher correlations across specifications are 0.88 in math and 0.91 in reading.

Based on these results for 5th grade, we expect that a hypothetical 4th-grade science VAM that controlled for students' science achievement in 3rd grade would produce estimates that are relatively highly correlated to the estimates that can be obtained currently. That some estimates are off the diagonal elements indicates that the presence or absence of a same-subject control affects the placement of some individual teachers in the distribution of effectiveness. In addition, the models with the same-subject baselines explain a greater portion of the overall variation in student scores. Specifically, the adjusted r-squared value is 3 percentage points higher in reading (an increase from 0.66 to 0.69) and 11 percentage points higher in math (an increase from 0.64 to 0.75) in the VAM specification that includes a same-subject control. On balance, we believe the evidence does not support discarding VAMs altogether when a same-subject baseline is unavailable, but presumably the accuracy and precision of the measures would improve if a same-subject prior score were available.

3. Controlling for Students' Prior Achievement Histories with a Fall or Spring Score

Because assessments are typically administered in the spring, most VAMs control for students' prior achievement histories using scores obtained at the end of the prior grade. Teacher effectiveness estimates using this approach therefore incorporate the effects of students' summer experiences, which can confound estimates of teacher contributions during the academic year. Measuring a student's achievement growth by testing at the beginning and near the end of the school year might produce a better attribution of learning to the teacher. But it introduces several new concerns as well, because schools would have to increase the time and resources devoted to testing and some teachers might deemphasize the fall assessment to produce larger gains.

Although there are concerns with both approaches, we sought to examine whether they nevertheless produce similar measures of teacher effectiveness given the currently available data.[27] In Table III.6, we compare teacher effectiveness quartiles generated from VAMs that differ by whether they control for math and reading 4Sight assessment scores using fall or spring scores. Because we can only access 4Sight scores in the pilot districts, the analysis is therefore limited to students in Allentown, Cornell, and Mohawk during the 2010–2011 year in grades 4 through 8. Each VAM specification included controls for PSSA math and reading scores from the prior grade but not from two prior grades, in addition to the math and reading 4Sight assessment scores. To maximize the teacher sample, we standardized the individual grade and subject effectiveness distributions and then combined teacher estimates across grades 4 through 8 as in Table III.4. We included teachers only if they had an effectiveness estimate under both specifications, although teachers are represented more than once if they teach students in math and reading.

[27] If teachers are given incentives to perform well based on these measures then the results might change.

Table III.6. Grade 4 Through 8 Math and Reading Teachers in Pilot Districts in Effectiveness Quartiles Based on Teacher VAMs with Fall or Spring Baselines Using PSSA Outcome Data

	Quartile of Effectiveness Based on Teacher VAM with Beginning- Year Fall 4Sight Baselines				
	1st (bottom)	2nd	3rd	4th (top)	Total
Quartile of Effectiveness Based on Teacher VAMs with Prior- Grade Spring 4Sight Baselines					
1st (bottom)	66 (77.6)	19 (22.4)	0 (0.0)	0 (0.0)	85 (100.0)
2nd	15 (17.6)	53 (62.4)	17 (20.0)	0 (0.0)	85 (100.0)
3rd	4 (4.7)	13 (15.3)	54 (63.5)	14 (16.5)	85 (100.0)
4th (top)	0 (0.0)	0 (0.0)	14 (16.5)	71 (83.5)	85 (100.0)
Total	85 (25.0)	85 (25.0)	85 (25.0)	85 (25.0)	340 (100.0)

Source: Mathematica calculations based on data from Pennsylvania and pilot districts' records.

Note: In each table cell, the first value is the number of teachers in the given cell and the second value (in parentheses) is the percentage of the row total that is represented by that cell.

PSSA = Pennsylvania System of School Assessment; VAM = value-added model.

The findings from this analysis support a conclusion that the timing of baseline controls might not make a large difference in the effectiveness measure for most teachers in these three districts. As in Table III.5, most of estimates are placed on the diagonal elements and nearly all the remaining estimates are off the diagonal by one quartile. The within-teacher correlations across specifications are 0.95 in math and 0.93 in reading. A comparison of the adjusted r-squared values by subject and grade across the two specifications indicates that using the fall baselines leads to a slight increase of about one percentage point in the percentage of variation in student scores that is explained by the model. On the whole, we find that fall and spring baselines produce similar teacher effectiveness estimates in Allentown, Cornell, and Mohawk, but that broader analyses should be conducted to support a general conclusion. In Phase 2 of the pilot, the expanded sample of districts should facilitate such an opportunity to conduct an analysis that is more representative of students in Pennsylvania.

D. Key Characteristics of Teacher Effectiveness Estimates Based on Pilot District Samples and Other Outcomes

More teachers can be included in value-added analyses by analyzing additional outcomes beyond PSSA scores in 4th through 8th grades. For teachers with students in grades 4 to 8, adding outcomes has the potential to make available multiple sources of information on each teacher's impact on his or her students. In Phase 1, we applied VAMs for measuring teacher effects to several additional outcomes using student samples from the Phase 1 districts in 2010–2011. First, we estimated teacher VAMs for non-PSSA outcomes in the lower elementary grades, in which the PSSA is not administered. Second, we estimated teacher VAMs based on PSSA assessments in 3rd and in 11th grades—grades that cannot be included in statewide analyses because a prior score is not available on a statewide basis. Thus, in all of the following analyses, the samples include only

students and teachers from the Phase 1 districts. This section describes key characteristics of the effectiveness estimates from these VAMs, which appear to differentiate among teachers except in 11th grade, the grade in which a larger sample of students is needed.

1. VAMs Based on Non-PSSA Assessments Administered by Pilot Districts

As we discussed in Chapter II, the pilot districts administer a number of assessments in lower elementary grades that are not covered by the PSSA. Adding these lower elementary grades to the analysis samples would substantially expand the set of teachers with value-added scores. We generated effectiveness estimates for teachers in the pilot districts based on the Progress Assessment (in Allentown) and the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) (in Allentown and Cornell). We analyzed the same key characteristics of these effectiveness estimates as we did for the PSSA-based estimates—namely, the extent of variation across teachers and the level of precision.

The effectiveness estimates appear to differentiate among teachers. In Table III.7, we show the estimates for teachers at the 15th and 85th percentiles. We omit the other percentiles given the smaller teacher sample sizes, but the patterns are similar to the PSSA distributions described earlier in the chapter. The 85th and 15th percentile teachers in Allentown differ in effectiveness by 9 percentage points on the writing Progress Assessment in 1st grade and by 11 percentage points on the math Progress Assessment in 2nd grade. By switching from the 15th to the 85th percentile teacher, a student originally at the median of these score distributions would be predicted to rise to the 73rd to 75th percentiles. There is less variation in teachers' impacts on 2nd-grade DIBELS scores, for which the 9-point difference in effectiveness between the 85th and 15th percentile teachers is equivalent to moving a student from the median to the 59th percentile of scores. Overall, these results (as well as results for additional assessments shown in Appendix Table C.4) generally suggest sizable variation in teachers' contributions to student scores.

Table III.7. Key Characteristics of Teacher Effectiveness Estimates Based on Selected Non-PSSA Tests Administered in the Pilot Districts

	Effectiveness of the Teacher at the Indicated Percentile Relative to the Effectiveness of the Average Teacher (in test scale points)		Percentage of Single-Cohort Teacher Effectiveness Estimates that Are Statistically Distinguishable from the Average
Outcome	15th	85th	
Progress Assessment, Writing, Grade 1 (Percentage Points)[a]	-5	4	38.0
Progress Assessment, Math, Grade 2 (Percentage Points)[a]	-5	6	34.8
DIBELS, ORF, Grade 2[b]	-5	4	18.2

Source: Mathematica calculations based on data from Pennsylvania and pilot districts' records.

Note: Findings are based on a 95 percent confidence interval and a one-cohort model with samples of teachers and students from the pilot districts in 2010-2011.

[a] Allentown only.
[b] Allentown and Cornell only.
ORF = oral reading fluency; PSSA = Pennsylvania System of School Assessment.

We caution, however, that analyzing the variation in teacher effects by the existing dispersion in student scores, as we have done in the preceding discussion, does not fully gauge whether this

variation is educationally meaningful. A closer examination of the content validity of these assessments—which is beyond the scope of our analysis—is necessary for determining whether this variation translates into substantive differences in students' skills and knowledge and, thus, is suitable for informing questions about educators' impacts on student performance.

The ability of the teacher VAM to make statistically reliable distinctions among teachers differs by outcome measure. For estimates based on the Progress Assessments, one-third of teachers can be statistically distinguished from average effectiveness. However, only 18 percent of teachers can be distinguished from the average based on their impacts on 2nd-grade DIBELS scores. Although the precision of the DIBELS-based estimates is not worse than that of the Progress-based estimates, there is less overall variation in teachers' impacts on 2nd-grade DIBELS scores. Therefore, a greater *proportion* of the variation in the DIBELS-based effectiveness estimates is due to random fluctuations and other sources of imprecision, making it more difficult to identify high- and low-performing teachers with high degrees of confidence. These differences in reliability across different types of effectiveness estimates could be factors in determining what weights PDE would like to place on these estimates in evaluating teachers.

2. VAMs Based on PSSA Assessments in 3rd and 11th Grades

We also expanded the grade-level coverage of teacher VAM estimates by using fall 4Sight scores as baseline achievement measures for PSSA outcomes in 3rd and 11th grades. In Table III.8, we show the characteristics of the resulting estimates.

In grade 3, as in grades 4 to 8, teachers' impacts on PSSA scores vary sizably. We find that the 15th and the 85th percentile teacher differ in effectiveness by 110 scale points in 3rd-grade math and by 55 scale points in 3rd-grade reading. Consistent with findings for higher grades on a statewide basis, 3rd-grade teacher impacts in Allentown, Cornell, and Mohawk are larger in math than in reading. Moreover, a larger percentage of teacher estimates can be distinguished from the average teacher in math than in reading, although the percentage in each subject is similar to those for one-cohort teacher VAMs statewide in grades 4 to 8. Overall, we interpret the results for the 3rd-grade VAMs as indicating that using fall 4Sight scores (or another measure, if available, that is aligned with PSSA content) as a baseline measure might be viable from an attribution standpoint. As indicated previously, there are potential resource and incentive-compatibility concerns involved with using a fall baseline in a teacher VAM that Pennsylvania policymakers should first consider carefully.

Table III.8. Key Characteristics of Teacher Effectiveness Estimates Based on PSSAs in 3rd and 11th Grades in the Pilot Districts

	Effectiveness of the Teacher at the Indicated Percentile Relative to the Effectiveness of the Average Teacher (in test scale points)		Percentage of Teacher Effectiveness Estimates that Are Statistically Distinguishable from the Average
Outcome	15th	85th	
PSSA, Math, Grade 3	-55	55	43.5
PSSA, Reading, Grade 3	-29	26	18.8
PSSA, Math, Grade 11	-21	17	3.4
PSSA, Reading, Grade 11	-17	12	0.0
PSSA, Writing, Grade 11	-19	24	0.0
PSSA, Science, Grade 11	-3	3	0.0

Source: Mathematica calculations based on data from Pennsylvania and pilot districts' records.

Note: Findings are based on a 95 percent confidence interval and a one-cohort model with samples of teachers and students from the pilot districts. The sample of teachers consists of those who served as teachers in 2010-2011 in Allentown, Cornell, or Mohawk school districts.

In contrast with the results from the 3rd-grade VAMs, the 11th-grade VAMs are not able to make reliable distinctions among teachers. The distributions of effectiveness for 11th-grade VAMS are more compressed than in lower grades, as indicated by a smaller difference between the effectiveness of the 15th and the 85th percentile teacher. The estimated teacher effects also have more imprecision than in the other models and, consequently, cannot distinguish even very high or low teacher contributions from the average.[28] The estimates are "shrunk," or pulled, more heavily toward the average to account for their greater imprecision (see Appendix A for a description of the shrinkage process). Part of the explanation for low precision is probably a lesser degree of alignment between the assessments and 11th-grade courses, which might be less likely to focus extensively on the skills measured in the PSSA. Keystone exams presumably will be more appropriate for teacher value-added use because they will be better aligned. As is, the results from Phase 1 for the 11th-grade VAMs would not be viable for use in an actual evaluation model. Precision might improve in Phase 2 when a much larger student sample can be included, but we would not expect substantial precision gains if the low precision is due to poor alignment between the content of the assessments and 11th-grade courses.

[28] We include students' 8th-grade PSSA scores as additional control variables in the 11th-grade teacher VAM along with fall 4Sight scores from grade 11. The inclusion of the 8th-grade scores, meant to enhance the controls for students' prior achievement histories, could nevertheless decrease precision if a large number of 11th-grade students are missing 8th-grade scores and thus have to be dropped from the analysis sample. We estimated an alternative specification of the 11th-grade VAM that omitted the 8th-grade scores. The alternative models included approximately 19 percent more students (that is, from about 715 to 850) but did not improve the ability of the VAM to distinguish between teachers and led to a reduction in the model r-squared value.

IV. RELATIONSHIPS BETWEEN TEACHER PRACTICES AND VALUE ADDED

In this chapter, we describe the analyses we conducted to examine relationships between value-added and teacher practices, as measured by the Phase 1 teacher observation rubric. We begin by summarizing the observation data obtained and the characteristics of the teachers who participated during Phase 1. Collectively, Phase 1 teachers are not dramatically different from teachers across Pennsylvania in terms of their demographics, master's degree attainment, and level of teaching experience. We then examine the variation in observation scores, which indicate that nearly all Phase 1 teachers were rated as either proficient or distinguished by their principals on the pilot rubric. Finally, we estimate the change in teacher value-added that is associated with a one-level increase in a teacher's score on different rubric components. Due to the small size of the pilot and a compressed distribution of observation scores, none of the resulting correlations are statistically significant, although most are numerically positive. We expect that the Phase 2 data will yield considerably more precision to these analyses.

A. The Phase 1 Teacher Observation Rubric and Score Distribution

The Pennsylvania teacher evaluation rubric administered to teachers in Phase 1 was based on the Framework for Teaching developed by Charlotte Danielson. The Danielson Framework includes 22 components grouped into four domains—planning and preparation, the classroom environment, instruction, and professional responsibilities. For Phase 1 of the Pennsylvania pilot, the stakeholder group focused on the 11 priority components in Table IV.1, consistent with similar work to improve teacher effectiveness in Pittsburgh Public Schools. [29]

Table IV.1. Danielson Framework Domains and Components, by Priority and Additional Components for the Pennsylvania Pilot

Priority Components		Additional Components	
Domain 1: Planning and Preparation			
1c:	Setting instructional outcomes	1a:	Demonstrating knowledge of content and pedagogy
1e:	Designing coherent instruction	1b:	Demonstrating knowledge of students
1f:	Designing assessment outcomes	1d:	Demonstrating knowledge of resources
Domain 2: The Classroom Environment			
2b:	Establishing a culture for learning	2a:	Creating an environment of respect and rapport
2d:	Managing student behavior	2c:	Managing classroom procedures
Domain 3: Instruction			
3b:	Using questioning and discussion techniques	3a:	Communicating with students
3c:	Engaging students in learning	3e:	Demonstrating flexibility and responsiveness
3d:	Using assessment in instruction		
Domain 4: Professional Responsibilities			
4a:	Reflecting on teaching and student learning	4d:	Participating in a professional community
4b:	System for managing students' data	4e:	Growing and developing professionally
4c:	Communicating with families	4f:	Showing professionalism

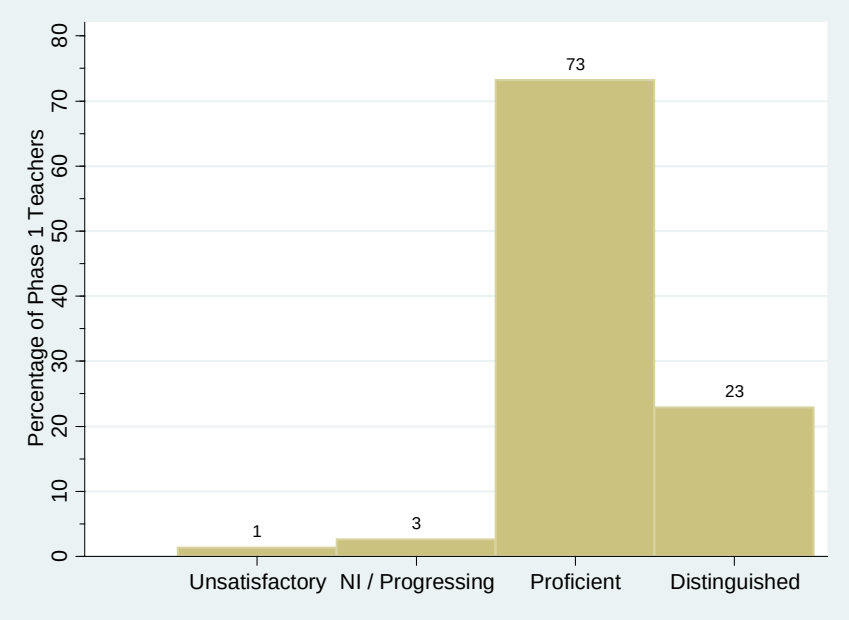
Source: Pennsylvania Teacher Evaluation Rubric from Phase 1.

[29] The Empowering Effective Teachers (EET) program is a joint project between Pittsburgh Public Schools and the Pittsburgh Federation of Teachers. Like Phase 1 of the Pennsylvania pilot, the EET program receives funding through the Bill & Melinda Gates Foundation.

Teachers were rated on a scale from one to four on each priority component (that is, 1 = Unsatisfactory; 2 = Needs Improvement for tenured teachers or Progressing for nontenured teachers; 3 = Proficient; and 4 = Distinguished). On the rubric, principals could also provide a short text description of the evidence on priority and additional components, although the narrative data are not used in this analysis. To help principals apply a consistent standard to their evaluations, the rubric included a component-by-component description of each level of performance. The evaluation matrix calculated an average rating across priority components and domains and a final rating that rounded the average rating to the nearest whole number.

In Figure IV.1, we show the distribution of final rating scores for the 153 Phase 1 teachers. Among Phase 1 teachers, 96 percent were rated as proficient or distinguished. One percent of these teachers were rated as unsatisfactory, which is the same percentage that the Pennsylvania Department of Education (PDE) found across the state for 2009–2010 under the existing observation protocol that differentiates only between satisfactory and unsatisfactory performance. Some principals rated every single teacher as proficient on all components. To the extent that more performance categories would be desirable for policy purposes, at the low end of performance the pilot rubric differentiated among teachers only slightly more than the current evaluation form does. The distribution of observation scores is more heavily concentrated in the proficient and distinguished categories than would be expected from applying typical results from the one-cohort value-added models (VAMs) at the precision levels reported in Table III.3. Specifically, approximately 30 percent of teachers in those models could be distinguished as above or below average. Depending on rubric rating definitions, we might therefore have expected a distribution in which about 15 percent of teachers are distinguished, 70 percent are proficient, and 15 percent are unsatisfactory or in need of improvement/progressing.

Figure IV.1. Distribution of Final Rating Scores for Phase 1 Teachers



Source: Observation data collected on Phase 1 teachers.

Notes: The data includes final rating scores on all 153 Phase 1 teachers from the four pilot districts.

NI = Needs Improvement.

The Consortium on Chicago School Research found a wider distribution of observation scores in a recent report documenting findings from a 2008 pilot study (Sartain et al. 2011). Like Pennsylvania's pilot, the Teacher Evaluation Pilot in Chicago implemented an observation protocol based on the Danielson Framework, although the two pilots might have adapted the Framework to their own needs. The distribution of scores, contrasted with Phase 1 scores in Table IV.2, indicates that principals in Chicago rated a larger percentage of teachers at the basic level than did principals in Phase 1, and a smaller percentage of teachers as proficient or distinguished.[30] Unlike the Pennsylvania pilot, classroom observations in the Chicago pilot included an external observer; observers were practitioners with extensive and ongoing training in the Danielson Framework for Teaching. Chicago's observing teachers were more conservative than principals in their ratings, with only 3 percent of teachers reaching the distinguished level.

Table IV.2. Final Ratings in Pennsylvania and Chicago, by Number and Percentage of Teachers

		Chicago Teachers	
Characteristic	Pilot Teachers in Phase 1 Rated by Principal	Rated by Principal	Rated by Observer
Distinguished	23	17	3
Proficient	73	53	67
Needs Improvement/Progressing (Basic in Chicago)	3	27	28
Unsatisfactory	1	3	2

Source: Ratings for Phase 1 teachers come from the pilot observation data. Ratings for Chicago teachers are reported in Sartain et al. (2011) Table 3.

Notes: The columns are based on data on 153, 4,747, and 4,852 ratings, respectively. Second observers were one of three individuals who were highly trained in the Danielson Framework for Teaching.

With data on only 153 Pennsylvania teachers, it is highly uncertain whether the distribution of scores obtained in Phase 1 is representative of scores that would be obtained by teachers across Pennsylvania in a larger pilot, especially because principals selected teachers for Phase 1 who had no prior history of unsatisfactory performance. However, we can glean from the data that Phase 1 teachers are not dramatically different from other Pennsylvania educators, at least in terms of broad demographic and professional characteristics. In Table IV.3, we compare teachers who participated in Phase 1 with all other educators in Pennsylvania based on their gender, race/ethnicity, educational attainment, and total years of experience.[31] The samples are not statistically different by gender and master's degree attainment. Relative to educators across Pennsylvania, a higher percentage of Phase 1 teachers were white and the distribution of total years of experience is more concentrated in the 6-to-10–years' category and less concentrated in the 21-or-more-years' category. Despite the statistical significance of the latter mean differences, the values are not dramatically different. Overall, we do not see any clear evidence to suggest that the distribution of observation scores would be

[30] The general description of basic in Chicago—understanding the components of teaching but implementing them sporadically—corresponds to the descriptions of needs improvement/progressing in the Phase 1 rubric.

[31] The comparison group includes other Pennsylvania educators rather than only Pennsylvania teachers because in the data we cannot differentiate teachers from other staff.

substantially different if a larger sample of teachers were observed, based on the teacher characteristics in Table IV.3.[32]

Table IV.3. Sample Characteristics of Nonpilot and Pilot Teachers

	Pilot Teachers in Phase 1 Districts	Nonpilot Educators in Pennsylvania	Statistically Significant Difference
Female	72	72	No
White	98	93	Yes
Master's Degree	46	51	No
Total Experience: 0-5 Years	20	25	No
Total Experience: 6-10 Years	34	23	Yes
Total Experience: 11-15 Years	20	19	No
Total Experience: 16-20 Years	12	12	No
Total Experience: 21 or More Years	14	21	Yes

Source: Mathematica calculations based on Pennsylvania data.

Note: The last column indicates statistically significant mean differences at the 5 percent level.

B. Observation Scores and Value- Added Scores for Phase 1 Teachers with VAM Estimates

As described in Chapter II, only 81 of the 153 teachers in the Phase 1 sample could be assigned value-added estimates. When only one assessment was available in a particular grade and subject, we applied estimates from that VAM to any respective Phase 1 teacher. When multiple assessments were available (as in 2nd-grade reading), we selected the VAM with the highest r-squared value.[33] If a teacher did not have a VAM estimate from the VAM with the highest r-squared among multiple assessments in a subject and grade, we used any VAM estimate that might be available for the teacher from the other assessments in the same subject and grade. This last step added only one additional teacher. All VAM estimates came from one-cohort models for 2010–2011 because the rubric covered teacher practices only in 2010–2011 and some outcomes were available in that year only.

There must be variation in observation scores to identify how changes in value added are associated with unit increases in observation ratings. However, the distribution of final ratings among these 81 teachers was even more skewed than in the overall sample. Thirty percent of these teachers received a distinguished rating and the remaining 70 percent received a proficient rating. None of these teachers had a final rating in the lower two categories. Given only two values for the final rating, we instead analyzed teachers' ratings on individual components and on their average score across priority components (that is, the final rating before rounding to the nearest whole number) because they are more continuous measures. We emphasize, however, that it is not clear

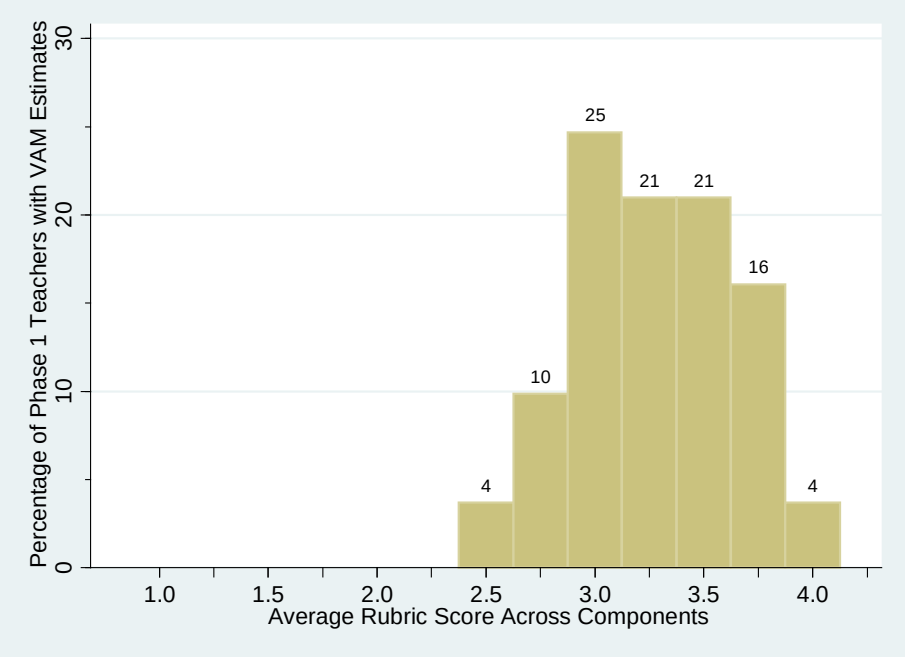
[32] For instance, the percentage of new teachers (defined as having 0 to 5 years of total experience) is not statistically different across groups. If the Phase 1 sample included a relatively low proportion of new teachers, we might expect a less skewed distribution for Pennsylvania overall because new teachers learn on the job during the first five years.

[33] This assessment was Dynamic Indicators of Basic Early Literacy Skills (DIBELS) – Phoneme Segmentation Fluency in 1st-grade reading and DIBELS – Oral Reading Fluency in 2nd-grade reading.

that a score of 2.75 or 3.25, for example, is meaningfully different than a score of 3.0, because the measurement properties of the observation rubric have not been examined.

In Figure IV.2, we depict the distribution of average ratings across priority components among Phase 1 teachers with VAM estimates. The average ratings are represented on the horizontal axis from 1.0 to 4.0, grouped into categories of 0.25 for illustrative purposes. The vertical axis indicates the percentage of teachers with a given average rating across priority components.

Figure IV.2. Distribution of Average Rating Scores for Phase 1 Teachers with VAM Estimates



Source: Observation data collected on 81 Phase 1 teachers with VAM estimates.

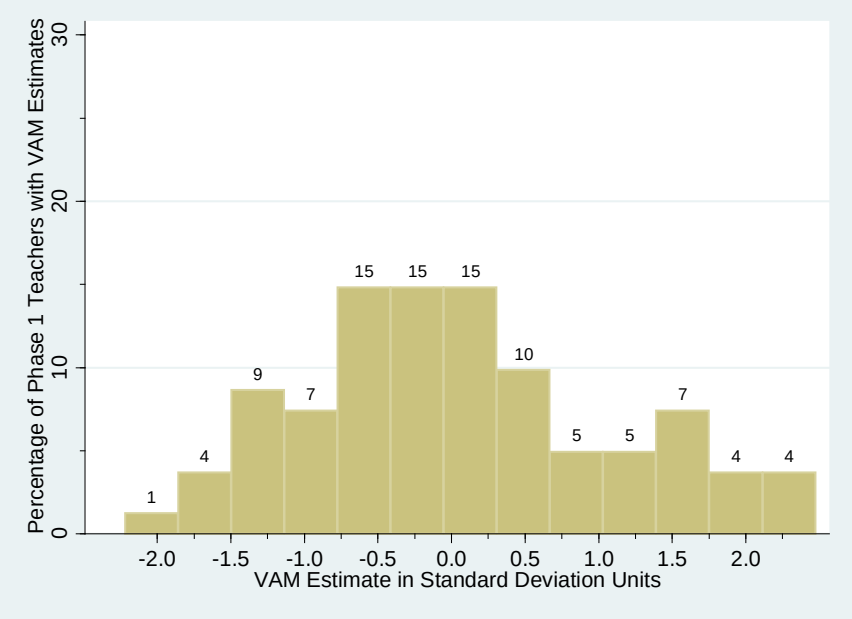
In Figure VI.3, we illustrate the range of VAM estimates for these same teachers.[34] Because the teacher VAM scores are not measured in rubric levels, we separated estimates into 13 equal-sized categories to correspond with the 13 rubric rating categories that are possible with a gradation of increments of 0.25 each.[35] The VAM distribution spreads out more than the distribution of rubric scores, with teachers placing in each of the 13 categories. The VAM distribution also looks more like a traditional bell curve centered on the average possible value, unlike rubric scores that are concentrated in the upper range of possible values. Given the compressed distribution of observation scores and the small sample, we do not expect to see any statistically significant relationships with value-added estimates unless very small differences in observation scores

[34] To obtain the VAM score distribution in Figure IV.3, we standardized individual estimates in Allentown and Mohawk (that is, the two districts in which Phase 1 teachers have VAM estimates) by grade and subject for each assessment before assigning them to pilot teachers. Thus, the VAM scores across teachers come from equivalent distributions.

[35] VAM estimates in Figure IV.3 are measured in standard deviation units. See Chapter III for a description of standard deviation units.

meaningfully differentiate between teachers in terms of their contributions to student achievement growth.

Figure IV.3. Distribution of VAM Scores for Phase 1 Teachers with VAM Estimates



Source: Mathematica calculations based on Pennsylvania data.

Note: Before collecting scores from across assessment VAMs, we standardized each estimate distribution to have a zero mean and a standard deviation of one among Allentown and Mohawk teachers.

C. Relationships Between Value- Added and Observation Scores

Using statistical models, we tested the relationships between teachers' estimated contributions to student learning and their observation scores for the 81 teachers with both types of effectiveness measures. The models compared the VAM score for individual teachers—as depicted in the prior figure—with their rubric ratings, holding constant average differences in teachers' VAM scores across districts, subjects, grade levels, and assessments. We conducted separate analyses for each priority component and domain-level average, and for the overall average rating across components.

The findings are interpreted as the predicted increase in teacher contributions to student learning from a one-level increase on the observation rubric. Expressing relationships between value-added and observation scores in this way is likely to be more informative for policymakers than as a correlation coefficient because the magnitude of the relationship is expressed in terms of student learning. Larger magnitudes indicate larger gains in student achievement for a one-level increase on a rubric component. As in a VAM, the statistical model can also indicate whether a particular relationship is statistically different from zero. By holding constant the variation across teachers in their districts, subjects, grade levels, and assessments, this model is likely to yield estimated relationships that are more accurate than a simple correlation coefficient. Because the VAM estimates (the outcome variable) have been standardized by grade and subject, comparisons of teachers can be legitimately made only within grade and subject. Therefore, it is necessary to force comparisons of the rubric score (the independent variable) to be made only within grade and subject.

In Table IV.4, we report the estimated relationships for all priority components and domain-level averages and for the overall average across priority components. None of the relationships are statistically significant, most likely due to the small sample size in Phase 1. The strongest relationships (even though not significant) are in the instruction domain, which is encouraging for future work. If a hypothetical estimate of 0.4 were statistically significant, it would mean that a one-level increase in a component score is associated with a 0.4 standard deviation increase in teacher effectiveness as measured by value-added.[36] An increase of this magnitude is equivalent to the additional contribution of a teacher at the 65th percentile of VAM scores above the contribution of a teacher at the 50th percentile. A student with median test scores would be expected to perform at the 53rd percentile if taught by the former rather than the latter teacher.

Table IV.4. Regression Coefficients Indicating the Standard Deviation Increase in Teacher Value-Added that Is Predicted for a One- Unit Increase in Rubric Scores

Domain or Component	Name	Estimate	Standard Error	Statistically Significant
Domain 1 Avg.	Planning and Preparation	0.04	0.29	No
1c	Setting instructional outcomes	0.13	0.27	No
1e	Designing coherent instruction	0.09	0.26	No
1f	Designing assessment outcomes	0.03	0.24	No
Domain 2 Avg.	The Classroom Environment	-0.06	0.28	No
2b	Establishing a culture for learning	-0.11	0.25	No
2d	Managing student behavior	0.03	0.30	No
Domain 3 Avg.	Instruction	0.44	0.32	No
3b	Using questioning and discussion techniques	0.39	0.25	No
3c	Engaging students in learning	0.35	0.30	No
3d	Using assessment in instruction	0.15	0.27	No
Domain 4 Avg.	Professional Responsibilities	0.01	0.31	No
4a	Reflecting on teaching and student learning	-0.07	0.34	No
4b	System for managing students' data	-0.29	0.26	No
4c	Communicating with families	0.24	0.25	No
Average Across Domains		0.13	0.36	No

Source: Mathematica calculations based on Pennsylvania student data and teacher observation data.

Note: Estimates are statistically significant at the 5 percent level if the absolute value of the estimate divided by its standard error is at least 1.96.

To explore these relationships further, we conducted similar analyses on partitioned samples by district, then by subject, and finally by grade range. The grade range partitions initially restricted the sample to teachers in grades 1 through 8 because the 11th-grade VAM estimates were highly imprecise. We then restricted the sample further to include only teachers with VAM estimates based on statewide samples (that is, PSSA outcomes in grades 4 through 8). With the exception of a few relationships for individual components in subject-specific analyses, none of these latter

[36] Specifically, this relationship is between rubric component scores and *estimated* value-added. As Jacob and Lefgren (2008) indicate, this relationship is likely to underestimate the relationship between rubric scores and a teacher's actual contributions to student learning because of measurement error in the value-added estimates. By applying the adjustment factor that they propose, we estimate that the actual relationships could be up to 23 percent larger than in Table IV.4. None of the qualitative inferences would change, however, because the fundamental problem is a lack of precision.

relationships were statistically significant. Given the large number of comparisons we conducted, we do not report these few relationships because we cannot rule out that they are statistically significant by chance. We also tried expressing the observation ratings as a series of indicator variables for final rating categories rather than as a continuous measure, but the results were similarly imprecise.

Although the findings from Phase 1 cannot differentiate among teacher practices in terms of their relationships with contributions to student learning, we do not view such analyses as hopeless for the future. Measuring these relationships requires a reliable measure in which evaluators are trained extensively to differentiate between levels of performance according to the rubric, and larger sample sizes. In short, ratings of professional practice have been found to be related to teachers' value-added. For instance, Tyler et al. (2010) studied the relationship between Cincinnati's Teacher Evaluation System (TES) and student achievement growth in math and reading. Like Pennsylvania's evaluation rubric, the TES is modeled on the Danielson Framework. The researchers found that a one-unit increase in the overall TES score was associated with student achievement gains that would move a student with the 50th percentile score to the 57th or 58th percentile, depending on the subject. Classroom management skills and, in reading, the use of inquiry-based teaching also were associated with greater gains in student achievement. In Chicago, nearly all of the Danielson Framework components had statistically significant relationships with value-added scores (Sartain et al. 2011). The teachers with the lowest rubric ratings tended also to have the lowest VAM estimates and vice versa. Milanowski et al. (2004) examined data from three school organizations that used rubrics based on the Danielson Framework: Cincinnati Public Schools, the Vaughn Next Century Learning Center charter school in Los Angeles, and the Washoe County school district in Nevada. A one-unit change in teacher evaluation scores on student achievement in these districts was similar in Cincinnati and Washoe to findings in Tyler et al. (2010) for math and reading and larger in the Vaughn charter school. Jacob and Lefgren (2008) concluded that principals in Chicago are typically able to distinguish between the teachers whose contributions to student achievement are the largest and the smallest, but are less able to distinguish teachers in the middle of the distribution. Finally, Rockoff and Speroni (2010) found that 3rd through 8th grade teachers in New York City who received higher ratings during their first year of teaching made greater contributions to student achievement in future years. The relationships they estimated remain intact even after additionally controlling for a teacher's value-added estimate from the first year of teaching. The authors also found evidence that observers vary in how they apply rating standards. Overall, the findings suggest that evaluation ratings based on both subjective and objective performance measures include more information than is conveyed by each measure independently, and that observation ratings standards should emphasize a high rate of interobserver agreement.

Finally, the Measures of Effective Teaching project sponsored by the Bill & Melinda Gates Foundation recently released a report that found positive relationships between student achievement gains and teacher practices using five classroom observation instruments including the Framework for Teaching (Kane and Staiger, 2012). The report was based on findings from 1,333 teachers in Charlotte-Mecklenburg, Denver, Hillsborough County (FL), New York City, and Memphis, who taught students in fourth through eighth grades in math and in English-language arts. Given these promising findings in the literature, we recommend pursuing these analyses further in Phase 2 of Pennsylvania's teacher evaluation pilot.

V. VALUE- ADDED RESULTS FOR PRINCIPALS

In this chapter, we analyze school-level value-added models (VAMs), considering whether they might produce valid, useful measures of principal effectiveness. A particular focus of our analyses is to assess the extent to which these VAMs disentangle the impacts of principals from the influences of other school-level factors beyond the principals' control.

In what follows, we begin by describing a method that, in theory, represents the best available approach to isolating principals' true impacts on student outcomes. We discuss how the limitations of this method prevent it from being applicable to real evaluations, and we present an alternative method, the school VAM, that is practicable but less able to separate principals' effects from the effects of other factors at their schools. The school VAM is the focus of all subsequent analyses. We assess the extent to which the school VAM produces estimates that approximate pure principal effects, and we conclude that this VAM should be regarded as estimating the effects of entire schools, which include both principal effects and the influence of other school-level factors. The final sections of this chapter describe several key characteristics of the effectiveness estimates generated by the school VAM.

A. An Approach to Estimating Pure Principal Effects

1. Challenges in Isolating Principals' Contributions to Student Achievement

A valid estimate of principals' effectiveness would isolate their effects on student achievement from the effects of other factors beyond the principals' control. Following the basic approach used for teacher VAMs, a natural starting point for assessing principals' effectiveness is to examine the average difference between actual and predicted outcomes among the students enrolled in a principal's school. This difference captures the contribution made by a principal's school to student achievement under the principal's tenure. In other words, *the starting point for estimating principal effectiveness is to estimate the effectiveness of the principal's school.*

The complication is that a school's effectiveness reflects more than just the effectiveness of the school's principal. It also reflects other school-specific characteristics and circumstances beyond the principal's control. First, preexisting teacher abilities—the abilities that teachers bring to the classroom regardless of the principal under whom they serve—contribute to school effectiveness. Principals affect student outcomes primarily by enabling their teachers to be more or less effective than expected, given their preexisting abilities. However, the mix of preexisting teacher abilities in a principal's school is often beyond his or her control. For example, a school located near a prestigious university might attract more highly motivated or capable teachers than a school in a less amenable location. The mix of abilities in a school's teaching staff can also reflect hiring decisions made by a principal's predecessor, and the current principal might have little flexibility to alter these decisions in the short run. Second, any differences among schools with respect to characteristics and

resources that are not accounted for in the VAMs—such as differences in funding, facilities, and neighborhood quality—can also lead to differences in the estimated effectiveness of the schools.[37]

Thus, a key analytic challenge of any statistical method that aims to identify the achievement effects of principals is to disentangle principals' true contributions to student outcomes—that is, pure principal effects—from the influence of other school-level factors. An estimation method used by some previous studies aims to address this challenge, which we discuss next.

2. Principal Transitions Model: Basic Structure

A type of VAM that we refer to as the *principal transitions model* provides an approach to distinguishing principal effects from the effects of other school characteristics and circumstances. Starting from estimates for schools' contributions to student outcomes, the model takes a further analytic step. It calculates how the same school's contribution differs under the leadership of different principals, and these differences serve to measure how effective a principal is relative to the other principals who have served at the same school. For example, if student outcomes relative to predicted outcomes rise when principal B succeeds principal A at a given school, then B is deemed to be more effective than A. Thus, the name of this model refers to the fact that only schools with leadership transitions during the considered period can be included in the analysis.

Because the principal transitions model is fundamentally based on comparing principals who have served at the same school, it controls for certain types of school-specific factors that are beyond principals' control. Specifically, the model controls for any school characteristics and circumstances that remain constant during the analysis period. These school-specific factors are common to all principals who have served at the same school, so they cannot contaminate comparisons of effectiveness among these principals. For example, if a school's proximity to a prestigious university does not change over time, then the resulting advantage in teacher recruitment will benefit equally all principals who have led this school and, thus, will not generate *differences* among the effectiveness estimates of these principals.

In the sparse research literature on the variation in and correlates of principal effectiveness, the principal transitions model is the most common type of principal VAM used by researchers. In fact, to our knowledge, all existing studies that have generated value-added estimates of principal effectiveness have used variants of the principal transitions model, either exclusively or in conjunction with alternative models (Branch et al. 2011; Dhuey and Smith 2011; Coelli and Green, forthcoming).[38] The popularity of this model stems from its ability to control for constant, school-specific influences on student achievement.

[37] The same types of school-level differences can be reflected in teacher value-added estimates as well, but the problem is less severe for teachers because the bulk of the variation in teacher effectiveness estimates is observed within—rather than between—schools (see Chapter III).

[38] Coelli and Green (forthcoming) augment the principal transitions model to allow a principal's impact on a school's effectiveness to evolve gradually over time with the principal's tenure at the school.

3. Limitations of the Principal Transitions Model

Although the principal transitions model might suit the research purposes of the previously described studies, it cannot be applied to real-world evaluations of principals. We identified and explored several limitations of this model and found them to be too severe to enable the model to be used in practice.

One major limitation of the principal transitions model is that it can generate effectiveness estimates for only a limited group of principals. Specifically, it can include only principals who have led schools in which a leadership transition has occurred during the analysis period. In a model with only one student cohort—that is, an analysis period of one year—the model is completely infeasible due to the lack of leadership transitions. Even over a three-year period, only a minority of schools undergo leadership transitions. For example, among all schools that contain students with 5th-grade Pennsylvania System of School Assessment (PSSA) math outcomes, 35 percent of schools (encompassing about 49 percent of principals) experienced at least one leadership transition during the 2008–2009 through 2010–2011 school years. In other words, excluding schools without leadership transitions would reduce by half the number of principals that can have effectiveness estimates. Thus, too many principals would have no effectiveness estimates if the principal transitions model were used for real evaluations.

For the principals who can have effectiveness estimates, the principal transitions model also limits the ways in which these principals can be compared on their performance. Comparisons can be made only within small connected networks of schools. Each connected network is a set of schools such that every member school has had at least one of its principals transfer to at least one other member school during the analysis period. By virtue of these transfers, principals from different schools within the same network can be compared; for example, if two principals from different schools are compared with a third principal who has served at both schools, they can, by implication, be compared with each other. However, the principal transitions model cannot determine how well a principal performed relative to another principal in a different network.

Connected networks are typically very small in the three-year period (2008–2009 through 2010–2011) covered by our analysis. Again, consider the set of principals from schools with 5th-grade PSSA math outcomes. Even among principals whose schools have undergone leadership transitions, 61 percent of these principals belong to networks with only a single school—their own school. In these networks, neither the predecessor(s) nor the successor(s) in the leadership of the school were observed in any other school with 5th-grade PSSA outcomes. Another 22 percent belong to networks with exactly two schools. Thus, only 17 percent of principals from schools with leadership transitions belong to connected networks with three or more schools, representing less than 9 percent of principals from all schools with or without leadership transitions. The key consequence is that the principal transitions model can determine how effective a principal is only relative to a very limited group of other principals. A meaningful evaluation system would need an assessment of a principal's effectiveness relative to a much broader comparison group.

Because the principal transitions model cannot be used in a real evaluation system, it is necessary to turn to an alternative value-added approach. We discuss this alternative approach next.

B. The School VAM as the Basis for Evaluating Principals

1. Analytic Approach

Given that the principal transitions model is infeasible for actual evaluations, the most straightforward alternative is simply to evaluate principals on the basis of their schools' contributions to student achievement. For each principal, this method—the school VAM—calculates the average effectiveness of the school(s) led by the principal during the analysis period. In other words, each principal receives a value-added score based on the difference between actual and predicted student outcomes averaged over all of the schools under his or her leadership during the analysis period.

The major limitation of using a school VAM to evaluate principals is that it bundles principals' true contributions with the effects of other school-level factors. Any types of differences across schools that are not accounted for in the model—such as differences in preexisting teacher abilities or school resources—could lead to differences in the value-added scores that principals receive. Compared with estimates from the principal transitions model, estimates from the school VAM have less validity as measures of pure principal effects.

However, the school VAM has several advantages. It does not suffer from the limitations of the principal transitions model; school value-added scores can be calculated for, and compared among, all eligible principals. This method also has the advantage of being conceptually straightforward: principals are held responsible for the extent to which their schools—including the teachers under their authority—affect student outcomes. Various districts and states, including Dallas and Tennessee, have implemented the approach of using school effectiveness measures to evaluate principals (see Lipscomb et al. 2010b). A school VAM has also been used in prior research to gauge the variation in principal effectiveness (Branch et al. 2011).

By virtue of being potentially applicable to a real evaluation system for principals, the school VAM is the central focus of the remainder of this chapter. Nonetheless, the following point is worth reiterating: *effectiveness estimates from school VAMs actually capture the effects of entire schools, including the effects of all instructional staff and educational inputs located at these schools.*

2. Comparison of School VAM and Principal Transitions Model

To interpret properly the effectiveness estimates from the school VAM, it is important to quantify the degree to which they deviate from pure principal effects. If these deviations were small, we could infer that school-level factors beyond the principals' control had only a small influence on school effectiveness, and we could regard estimates from the school VAM as being primarily indicative of principals' contributions. Large deviations, on the other hand, would suggest these estimates were poor measures of principal effects.

To carry out this analysis, we use estimates from the principal transitions model as benchmarks with which estimates from the school VAM are compared. Despite being inapplicable to real evaluations, estimates from the transitions model nevertheless represent our best estimates of pure principal effects and, as such, can serve as a useful point of comparison. Thus, to the extent that the two models yield more similar results, there will be greater justification for interpreting estimates from the school VAM as primarily reflecting principals' true contributions to student achievement.

Using Pennsylvania data, we obtained effectiveness estimates from the school VAM and the principal transitions model for the set of principals that can be included in both models. Because the initial estimates from the school VAM compare principals from all schools—not just the principals in the same connected network, as in the principal transitions model—we first converted these estimates so that they would have the same meaning as those from the transitions model. From each principal's effectiveness estimate, we subtracted the average effectiveness estimate in the principal's network. As a result, the final estimates from both models capture the deviation of a principal's effectiveness from the average in the same network. For each of the two models, we placed principals into quartiles based on how much they outperform or underperform the average principal in their network.

Table V.1 compares the quartiles into which principals are placed based on the two models. This table addresses the question: *To what extent do the school VAM and principal transitions model rank principals similarly on their measured effectiveness?* Each row of the table represents a particular quartile of principals from the transitions model, and row entries show the number and percentage of those principals who are placed into each of the four quartiles based on the school VAM. The diagonal entries of the table represent the cases in which the two models coincide in placing principals into the same effectiveness quartile. For ease of presentation, we show results based on two outcomes—5th-grade PSSA math scores and 8th-grade PSSA reading scores—and pool the analyses related to both outcomes together into a single table.

We find a moderate degree of consistency between the effectiveness rankings produced by the two models. As shown by the diagonal (upper left to lower right) entries of Table V.1, about half of the principals in the analysis are placed into identical quartiles by the two models. Moreover, for most principals, their effectiveness estimates based on the school VAM differ by no more than one quartile from their effectiveness estimates based on the principal transitions model. The simple correlations between effectiveness estimates from the two models—0.39 in 5th-grade math and 0.58 in 8th-grade reading—also yield the same conclusion that the two models are moderately consistent with each other.

Although these results are encouraging, a noticeable minority of principals still receive a ranking from the school VAM that is substantially different from their transitions model ranking. For example, of the principals in the bottom quartile identified by the transitions model, 27 percent are in the top two quartiles identified by the school VAM. Similarly, 24 percent of principals in the top quartile from the principal transitions model are in the bottom two quartiles from the school VAM. Notably, there is less consistency between the school VAM and the principal transitions model than there is between the various teacher VAMs from Chapter III that used different baseline achievement controls (see Tables III.5 and III.6). The quartiles of effectiveness into which teachers were placed by these different VAMs rarely differed by more than one quartile.

Table V.1. Counts and Percentages of Principals in Effectiveness Quartiles Based on Principal Transitions Model and School VAM

	Quartile of Effectiveness Based on School VAM				
	1st (bottom)	2nd	3rd	4th (top)	Total
Quartile of Effectiveness Based on Principal Transitions Model					
1st (bottom)	69 (53.5)	25 (19.4)	15 (11.6)	20 (15.5)	129 (100.0)
2nd	31 (23.7)	63 (48.1)	26 (19.8)	11 (8.4)	131 (100.0)
3rd	13 (10.4)	25 (20.0)	58 (46.4)	29 (23.2)	125 (100.0)
4th (top)	16 (12.6)	15 (11.8)	29 (22.8)	67 (52.8)	127 (100.0)
Total	129 (25.2)	128 (25.0)	128 (25.0)	127 (24.8)	512 (100.0)

Source: Mathematica calculations based on Pennsylvania data.

Note: In each table cell, the first value is the number of principals in the given cell and the second value (in parentheses) is the percentage of the row total that is represented by that cell. Findings are based on statewide samples of principals with effectiveness estimates in either grade 5 math or grade 8 reading; principals with both types of effectiveness estimates are counted twice. The sample of principals consists of those who served as principals at any time from 2008-2009 to 2010-2011. Only principals with effectiveness estimates from both the school VAM and the principal transitions model are included in the analysis. To construct this table, principals were placed into cells separately for each of the two examined outcomes, and the resulting counts in each cell were aggregated across the two outcomes.

Overall, the results from Table V.1 suggest that *estimates from the school VAM are an informative but imperfect measure of principals' contributions to student learning*. Given the moderate consistency of these estimates with those from the transitions model, some of the variation in these estimates among principals is likely to capture true differences in principal quality. However, the discrepancies with the transitions model also suggest that some of the variation in these estimates is picking up school-level differences outside of the principals' control. Based on this evidence, we continue to believe that the school VAM estimates measure the contributions of entire schools to student achievement, and should not be described as principal value-added measures.

C. Key Characteristics of School Effectiveness Estimates Based on PSSA Outcomes

We proceed to describe several empirical features of the effectiveness estimates from the school VAM, including the amount of variation and the extent of statistical uncertainty in these estimates. All of these features influence the extent to which the VAM can distinguish effective and ineffective schools.

For ease of exposition, we will refer to schools as the entities being compared when describing school VAM estimates. As described previously, the method actually generates one estimate per principal based on the effectiveness of the principal's school(s) under his or her tenure; a principal who has led multiple schools is assigned a single estimate based on the average effectiveness of those schools. Nevertheless, referring to schools as the units of comparison emphasizes, once again,

that the school VAM fundamentally measures the contributions of entire schools—not only principals—to student outcomes.

This section focuses on school VAM estimates for which PSSA scores are the outcomes of interest. As discussed in Chapter II, VAMs based on PSSA outcomes yield effectiveness estimates for schools across the entire state, enabling the analysis of VAM characteristics to be supported by large samples. Unless otherwise noted, most of the following analyses pertain to VAM estimates that use data from three student cohorts.

1. Variation in Measured Effectiveness Across Schools

In order for VAMs to be informative in distinguishing effective and ineffective schools, it is necessary that there exist meaningful variation in the effectiveness estimates across schools. To document this variation, we calculated the extent to which schools at selected percentiles differ from the average school. Table V.2 presents these measured differences for selected grade-subject combinations. In addition, Appendix Table C.5 expresses differences in effectiveness estimates in terms of standard deviations of student scores for all grade–subject combinations.

Table V.2. Distribution of School Effectiveness Estimates for Selected PSSA Outcomes

	Effectiveness of the School at the Indicated Percentile Relative to the Effectiveness of the Average School (in PSSA scale points)					
Outcome	5th	15th	25th	75th	85th	95th
Math PSSA, Grade 5	-67	-40	-26	24	43	69
Math PSSA, Grade 11[a]	-94	-50	-33	35	53	85
Reading PSSA, Grade 8	-45	-28	-18	19	29	46
Science PSSA, Grade 4	-68	-43	-26	27	43	69

Source: Mathematica calculations based on Pennsylvania data.

Note: Unless otherwise noted, findings are based on a three-cohort model with statewide samples of schools, principals, and students. The sample of principals consists of those who served as principals in every year from 2008-2009 to 2010-2011.

[a] Findings are based on a two-cohort model because three years must elapse between the baseline and outcome scores.

The results indicate that effectiveness estimates vary considerably across schools. For example, by attending the 15th percentile school, a 5th-grade student would score 40 scale points lower on the math PSSA than he or she would score by attending the average school. On the other hand, this student would score 43 scale points higher in the 85th percentile school than in the average school. Thus, the 85th and 15th percentile schools differ in their effectiveness by 83 PSSA points. This scale point difference can also be interpreted with reference to the statewide distribution of student test scores. By switching from the 15th to the 85th percentile school, a 5th-grade student originally at the median of the statewide distribution of math scores would be predicted to rise to the 65th percentile.[39]

[39] To make this calculation, we divided 83 scale points by 223 scale points, the standard deviation of 5th-grade PSSA math scores shown in Appendix Table C.5. Thus, an 83 scale point difference amounts to a difference of 0.37

Sizable variation in the school effectiveness estimates is also observed for other grade–subject combinations. The 85th and 15th percentile schools differ in their estimated effects on PSSA scores by 103 points in 11th-grade math, 57 points in 8th-grade reading, and 86 points in 4th-grade science. As we found for teachers in Chapter III, variation in schools' effects appears to be smaller in reading than in math for every grade level (see Appendix Table C.5).

Our results for the variation in school effectiveness are in the range of the variation found by prior research. For example, for math PSSA outcomes, we find that the difference in effectiveness between the 85th and 50th percentile schools, expressed in standard deviations of student scores, ranges from 0.15 to 0.21 across grades (see Appendix Table C.5). Similarly, using math test scores from Texas in grades 3 through 8, Branch et al. (2011) found a corresponding difference of 0.21 standard deviations.

Notably, the variation in school effectiveness estimates is similar in magnitude to the variation in teacher effectiveness estimates described in Chapter III. The reason that effectiveness estimates do not vary considerably more for teachers than for schools is that VAMs account for imprecision in both types of estimates by "shrinking," or pulling, these estimates toward their respective averages. Teachers' estimates are pulled more heavily toward the average due to their greater imprecision. Appendix A describes this shrinkage approach in further detail.

In summary, there is meaningful variation in schools' estimated impacts on PSSA scores. As a consequence, if an evaluation system for principals used the school VAM, it would be feasible for the system to delineate groups of principals that differed in their school performance estimates by a substantively important magnitude. This does not mean, however, that school VAM estimates are *valid* measures of principal performance, because they also include aspects of school performance that are outside of principals' control.

2. Statistical Uncertainty in the School Effectiveness Estimates

As with the teacher effectiveness estimates examined in Chapter III, the effectiveness estimates for schools contain some degree of statistical uncertainty. To the extent that there is less uncertainty in these estimates, chance errors in estimation—due, for instance, to random fluctuations in the composition of a school's students—exert less influence on these estimates. Thus, quantifying statistical uncertainty is important for determining which of the measured differences across schools are unlikely to have arisen purely by chance.

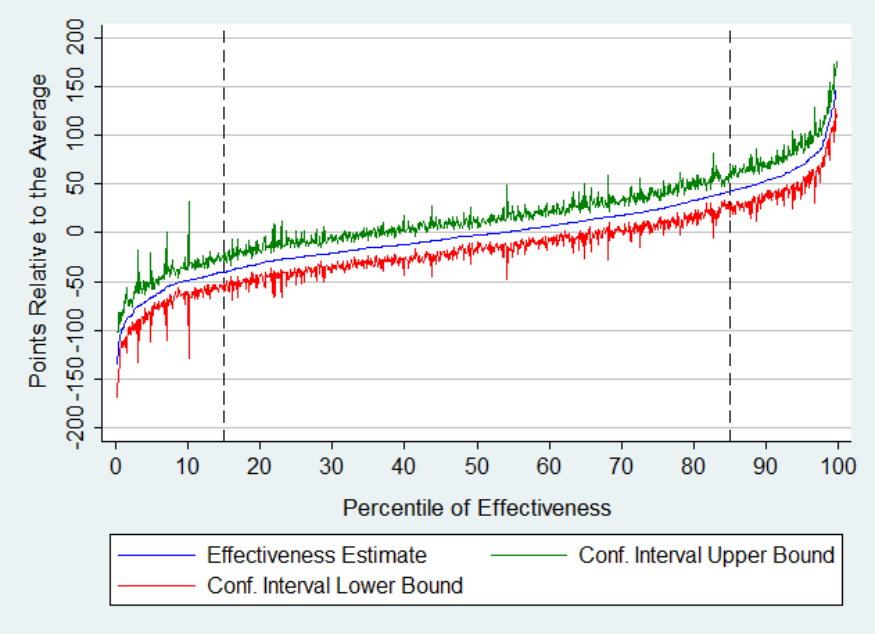
Figure V.1 provides a visual depiction of both the statistical uncertainty in the school effectiveness estimates and the variation in these estimates across schools based on 5th-grade PSSA math outcomes. As with similar figures for teachers in Chapter III, this figure plots (on the vertical axis) the school effectiveness estimates, as well as the lower and upper bounds of the 95 percent confidence intervals for these estimates, against the school's percentile rank (on the horizontal axis). Estimates whose confidence intervals lie completely above or completely below zero—defined to be

(continued)

standard deviations in the distribution of student scores. In the assumed normal distribution for student scores, moving from the 50th to the 65th percentile is equivalent to an increment of 0.37 standard deviations.

the effectiveness of the average school—are statistically distinguishable from average school effectiveness.

Figure V.1. Distribution of School Effectiveness Estimates and 95 Percent Confidence Intervals of School Effectiveness Estimates for Math PSSA Grade 5 Scores



Source: Mathematica calculations based on Pennsylvania data.

Note: Findings are based on a three-cohort model with statewide samples of schools, principals, and students. The sample of principals consists of those who served as principals in every year from 2008-2009 to 2010-2011. In the figure, vertical dotted lines are drawn at the 15th and 85th percentiles.

As shown in Figure V.1, nearly all of the schools in the top and bottom quartiles of the performance distribution have effectiveness estimates that are statistically distinguishable from average effectiveness. For example, both the 15th and 85th percentile schools, represented by dotted lines in Figure V.1, differ from average effectiveness by a statistically significant extent. Naturally, the share of schools that is statistically different from the average declines as the estimates move closer to the middle of the performance distribution. Effectiveness estimates between the 40th and 60th percentiles are generally not statistically different from average effectiveness. The figure thus provides a visual indication that the VAM can identify highly effective and highly ineffective schools but is less able to make distinctions among schools near the middle of the performance distribution.

To summarize the extent of statistical uncertainty in the school effectiveness estimates, Table V.3 provides the number and share of schools whose effectiveness estimates are statistically different from average effectiveness. For the three-cohort models—the main focus of our analysis—the share of schools that can be distinguished from the average ranges from 58 to 69 percent, depending on the outcome measure. Comparing these results with those from Table III.3, we find that larger fractions of schools than teachers have effectiveness estimates that are statistically distinguishable from the average. The reason is that schools' effectiveness estimates are typically based on larger samples of students and, hence, have greater precision.

Table V.3. Number of Schools with Effectiveness Estimates Reported and Share of Reported Estimates that Are Statistically Different from the Average, by Number of Cohorts Used in Estimation

	Number of Principals		Percentage of School Effectiveness Estimates that Are Statistically Distinguishable from the Average	
Outcome	1-Cohort Model	3-Cohort Model	1-Cohort Model	3-Cohort Model
Math PSSA, Grade 5	1,336	1,079	53.0	66.4
Math PSSA, Grade 11	626	577[a]	59.7	68.3[a]
Reading PSSA, Grade 8	755	580	41.2	58.3
Science PSSA, Grade 4	1,427	1,166	54.1	69.0

Source: Mathematica calculations based on Pennsylvania data.

Note: Findings are based on a 95 percent confidence interval and statewide samples of schools, principals, and students. In the one-cohort model, the sample of principals consists of those who served as principals in 2010-2011. Unless otherwise noted, the sample of principals in the three-cohort model consists of those who served as principals in every year from 2008-2009 to 2010-2011.

[a] Findings are based on a two-cohort model because three years must elapse between the baseline and outcome scores.

Given that statistical imprecision is a less severe problem for estimating school effectiveness than for estimating teacher effectiveness, an issue for consideration by the Pennsylvania Department of Education (PDE) is whether to use fewer than three cohorts for the school VAM. A key advantage of using a smaller number of (the most recent) cohorts is that the VAM provides a more up-to-date measure of school performance. This advantage must be weighed against the decrease in the precision of the estimates. For the outcomes shown in Table V.3, the fractions of schools that are statistically distinguishable from the average are lower in one-cohort models than in three-cohort models by 9 to 17 percentage points. For example, in a school VAM based on 5th-grade PSSA math scores, 53 percent of schools are statistically different from the average in a one-cohort model, whereas the corresponding percentage is 66 percent in a three-cohort model. In other words, it is more difficult for one-cohort models to distinguish true performance differences among schools from random fluctuations in the outcomes of their students. Of note, however, is that the shares of schools that are statistically distinguishable from the average in one-cohort models are at least as high as the corresponding shares for teachers in *three-cohort* models (see Tables III.3 and V.3). Thus, the decrease in precision from using fewer cohorts could be more tolerable in the case of schools than in the case of teachers.

D. Key Characteristics of School Effectiveness Estimates Based on Outcomes Other than PSSA Scores

Analyzing additional outcomes beyond PSSA scores has the potential to provide a more comprehensive picture of each school's impact. In Phase 1, we explored two broad categories of additional outcomes to which we applied school VAMs. First, as we did for teachers in Chapter III, we estimated school VAMs based on assessments administered by pilot districts to the lower elementary grades—grades excluded from the PSSA-based models. Second, we examined schools' impacts on key nonassessment outcomes—holding power and attendance—that are regarded as important precursors of academic success. This section describes key characteristics of the effectiveness estimates from these VAMs.

1. VAMs Based on Assessments Administered by Pilot Districts

Given that the PSSA does not cover all grade–subject combinations, the use of data from district assessments can expand the set of grades and subjects that can be included in school VAMs. As we discussed in previous chapters, the pilot districts administer a number of assessments at lower elementary grades not covered by the PSSA. However, expanding coverage of VAMs to the lower elementary grades yields a different type of benefit for principal evaluation systems than for teacher evaluation systems. Although adding lower elementary grades to the analysis samples would substantially expand the set of teachers with value-added scores, it would lead to only modest increases in the number of principals with school value-added scores. Most schools with lower elementary grades also contain upper elementary grades covered by the PSSA, enabling their principals already to have school VAM estimates based on these PSSA outcomes. For example, in a three-cohort model, whereas 1,249 principals would have a school VAM estimate from at least one 4th- or 5th-grade PSSA assessment, the inclusion of all remaining elementary grades (K–3) into the VAMs would bring, at most, 51 additional principals into the analyses—a 4 percent increase. Thus, for principal evaluations, the primary benefit of applying VAMs to the lower elementary grades is to be able to measure elementary schools' effectiveness based on the widest possible set of grades.

To assess the potential for including district assessments in school VAMs, we generated effectiveness estimates for schools in the pilot districts based on locally administered assessments (see Chapter II for a discussion of the assessments and samples). We analyzed the same key characteristics of these effectiveness estimates as we did for the PSSA-based estimates—namely, the extent of variation across schools and the level of precision.

Schools in the pilot districts appear to differ in their impacts on district assessment scores by meaningful magnitudes. Table V.4 shows the effectiveness estimates (relative to average effectiveness) for the schools at the 15th and 85th percentiles; we omit other percentiles to maintain participants' confidentiality, given the small sample sizes in this analysis. The 15th and 85th percentile schools in Allentown differ in effectiveness by 14 percentage points on the Writing Progress Assessment in 1st grade and by 9 percentage points on the Math Progress Assessment in 2nd grade. To interpret these differences, it is again instructive to convert them to increments within the distribution of student test scores. A student who would have had the median Progress score in Allentown if assigned to the 15th percentile school would, instead, be at the 71st to 84th percentiles of Progress scores if assigned to the 85th percentile school. There is less variation in schools' impacts on 2nd-grade Dynamic Indicators of Basic Early Literacy Skills (DIBELS) scores, even though the sample includes two districts (Allentown and Cornell). In the distribution of DIBELS scores, the eight-point difference in effectiveness between the 15th and 85th percentile schools is equivalent to moving a student from the median to the 59th percentile of scores. Overall, however, these results (as well as results for additional assessments shown in Appendix Table C.6) generally suggest sizable variation in schools' contributions to student scores on the pilot districts' assessments.

Table V.4. Key Characteristics of School Effectiveness Estimates Based on Selected Tests Administered in the Pilot Districts

	Effectiveness of the School at the Indicated Percentile Relative to the Effectiveness of the Average School (in test scale points)		Percentage of School Effectiveness Estimates that Are Statistically Distinguishable from the Average
Outcome	15th	85th	
Progress Assessment, Writing, Grade 1 (Percentage Points)[a]	-8	6	53.8
Progress Assessment, Math, Grade 2 (Percentage Points)[a]	-4	5	66.7
DIBELS, ORF, Grade 2[b]	-4	4	13.3
PSSA, Math, Grade 3[c]	-35	31	37.5
PSSA, Reading, Grade 3[c]	-26	25	37.5

Source: Mathematica calculations based on Pennsylvania data.

Note: Findings are based on a 95 percent confidence interval and a one-cohort model with samples of schools, principals, and students from the pilot districts. The sample of principals consists of those who served as principals in 2010-2011.

[a] Allentown only.
[b] Allentown and Cornell only.
[c] Allentown, Cornell, and Mohawk only.
DIBELS = Dynamic Indicators of Basic Early Literacy Skills; ORF = oral reading fluency.

As indicated in Chapter III, we caution that expressing the variation in school effects in terms of the existing dispersion in student scores does not fully gauge whether this variation is educationally meaningful. A closer examination of the content validity of these assessments is necessary for determining whether this variation translates into substantive differences in students' skills and knowledge.

The ability of the school VAM to make statistically reliable distinctions among schools differs by outcome measure. As shown in the last column of Table V.4, for estimates based on the Progress Assessments, one-half to two-thirds of schools can be statistically distinguished from average effectiveness. However, only 13 percent of schools can be distinguished from the average based on their impacts on second-grade DIBELS scores. As we described in Chapter III, a greater proportion of the variation in the DIBELS-based effectiveness estimates is due to random fluctuations and other sources of imprecision, making it more difficult to identify high- and low-performing schools with high degrees of confidence.

In addition to using outcomes from non-PSSA assessments as a means of expanding grade-level coverage, we also explored applying school VAMs to 3rd-grade PSSA outcomes by using 4Sight scores as baseline achievement measures. The final two rows of Table V.4 show the characteristics of the resulting estimates. Consistent with schools' impacts on PSSA scores at other grade levels, impacts on 3rd-grade PSSA scores vary substantially across schools; the 85th and 15th percentile schools differ in effectiveness by 66 PSSA points in math and 51 PSSA points in reading. A slightly lower proportion (38 percent) of schools can be statistically distinguished from the average school on the basis of the 3rd-grade PSSA VAMs than on the basis of the other one-cohort PSSA VAMs (shown previously in Table V.3). One reason is that the baseline scores and other student characteristics do not explain as much of the outcome variance in the 3rd-grade PSSA VAMs as in

most of the other PSSA VAMs that we estimated.[40] However, it is unclear whether the pilot districts are unique in this respect, or whether it is generally the case statewide that fall 4Sight scores have less power than prior PSSA scores to predict current-year PSSA scores. Later pilot phases, with a larger sample of districts, can address this question more definitively.

The results for the 3rd-grade VAMs indicate that using 4Sight scores as baseline measures is one potentially feasible strategy for including 3rd grade in the school VAMs. The benefits of including 3rd grade through this strategy should be weighed against several additional considerations. First, there are likely to be costs of introducing the 4Sight assessment into districts that currently do not use it. Second, like any other fall assessment, the fall 4Sight would be used only as a baseline measure and not as an outcome measure in any VAM, which might give educators an incentive to deemphasize—or even intentionally depress performance on—this assessment. Third, the fall 4Sight is not the only baseline measure that could potentially be used in a 3rd-grade VAM, but it is likely to be the one most closely aligned with the content of the 3rd-grade PSSA.

2. VAMs Based on Nonassessment Outcomes

Outcomes for VAMs do not necessarily have to be limited to test scores. If using multiple measures of student outcomes in evaluations is deemed to be a priority, then schools' impacts on nonassessment outcomes merit consideration. In fact, for various types of nonassessment outcomes, estimating the impacts of entire schools is more feasible than estimating the impacts of individual teachers. For example, a student's persistence in remaining enrolled in high school—which we call holding power—and a student's attendance rate are likely to be affected by multiple teachers who instruct this student, as well as by a myriad of other factors at the student's school, such as school culture. Although the influence of all of these school-based factors can be bundled together into an estimate of the school's impact, it would be much more difficult to isolate the effects of individual teachers on these outcomes.

We estimated school VAMs for the two nonassessment outcomes mentioned previously—holding power and attendance rate—because they are regarded as important intermediate outcomes that feed into educational attainment and achievement outcomes. A school's effectiveness estimate for a nonassessment outcome has a similar meaning as it does for an assessment outcome: it is the difference between what the school's students achieve and what they would have achieved if they had been assigned to the average school.

Table V.5 shows the distribution of school effectiveness estimates for nonassessment outcomes. Because the holding power VAMs are based on large, statewide samples and the attendance rate VAMs are based on the pilot districts only, we show several more percentiles of the school effectiveness distribution for holding power than for attendance.

[40] As shown in Appendix C.6, the R-squared values from the 3rd-grade PSSA VAMs range from 0.64 to 0.69. However, the R-squared values (not shown in the tables) for the one-cohort models based on 5th-grade math PSSA and 8th-grade reading PSSA scores are 0.77 and 0.74, respectively.

Table V.5. Key Characteristics of School Effectiveness Estimates Based on Nonassessment Outcomes

	Effectiveness of the School at the Indicated Percentile Relative to the Effectiveness of the Average School (in Percentage Points)						
Outcome	5th	15th	25th	75th	85th	95th	Percentage of Effectiveness Estimates that Are Statistically Distinguishable from the Average
Holding Power, Grade 9	-30.8	-21.3	2.4	7.0	7.4	10.6	92.2
Holding Power, Grade 10	-32.5	-21.2	2.8	6.8	7.3	11.0	92.0
Holding Power, Grade 11	-81.5	1.4	7.0	10.7	11.3	20.6	90.0
Attendance Rate, Grades 4-12	--[a]	-0.5	--[a]	--[a]	0.6	--[a]	13.8

Source: Mathematica calculations based on Pennsylvania data.

Note: Findings on holding power in grades 9 and 10 are based on a two-cohort model with statewide samples of schools, principals, and students; for these outcomes, the sample of principals consists of those who served as principals in both 2009-2010 and 2010-2011. Findings on holding power in grade 11 are based on a one-cohort model with statewide samples of schools, principals, and students. Findings on attendance are based on a one-cohort model with samples of schools, principals, and students from Allentown, Mohawk, and Northwest Tri-County. For the one-cohort models, the sample of principals consists of those who served as principals in 2010-2011. All analyses are based on a 95 percent confidence interval.

[a] Due to the small sample sizes in this analysis, effectiveness estimates at these percentiles are suppressed in order to protect the confidentiality of the sample members.

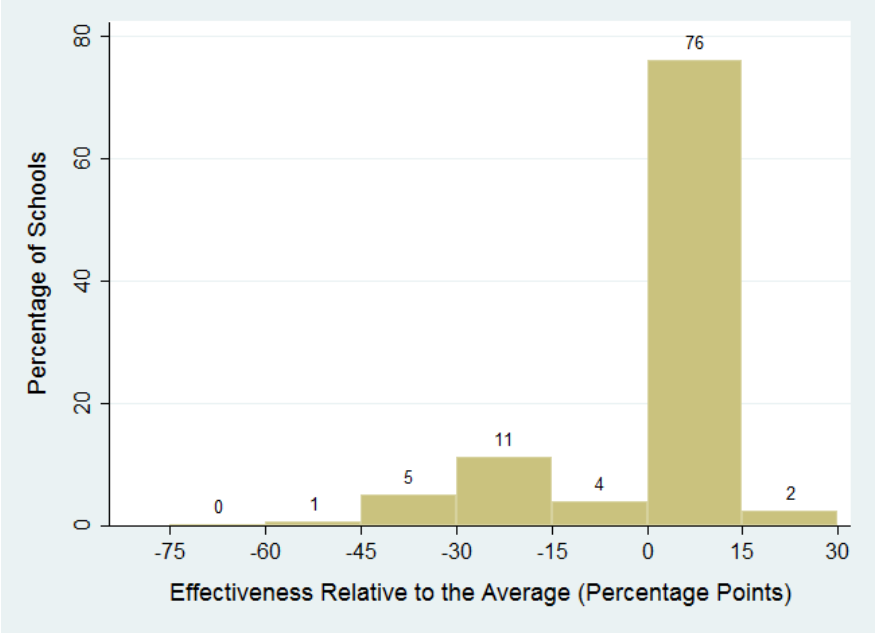
There are striking differences among schools in their effectiveness at keeping students enrolled in high school. A 9th-grader who attends the 85th percentile school is 29 percentage points more likely to enroll in a Pennsylvania public school again in the following year than if he or she had attended the 15th percentile school. Large differences are also observed at the other high school grades.

The distribution of schools' effects on holding power is highly skewed, as shown in both Table V.5 and the histogram in Figure V.2. The bottom 6 percent of schools have extremely negative impacts on holding power, lowering their students' probability of staying enrolled by more than 30 percentage points relative to the average school. These schools pull the average effectiveness in the sample downward by such a magnitude that a large majority (78.5 percent) of schools have effectiveness estimates that are above average, many of which are statistically significantly above average. Moreover, nearly all of these above-average schools have effectiveness estimates that are within 15 percentage points of one another. These patterns suggest that there are stark differences between the worst-performing high schools and the other schools in the state with respect to their impacts on keeping students enrolled.

On the basis of these results, school effectiveness estimates for holding power appear to be an informative tool for identifying high schools that perform poorly in keeping their students enrolled in Pennsylvania's public schools. However, a number of factors merit attention when interpreting these estimates. First, these estimates are only as valid as the underlying data on student enrollment. To the extent that any students' enrollment records are missing from the Pennsylvania Information Management System (PIMS), the schools that these students attended in the previous year will be erroneously regarded as having failed to "hold on" to these students. Second, holding power in

grades 9 through 11 is an intermediate outcome for a final outcome that is usually of greater interest—graduation. We could not estimate schools' impacts on graduation (and/or holding power in grade 12) in Phase 1 because there were insufficient years of data to control for the 8th-grade PSSA scores of 12th-graders.[41] With an additional year of student data, it will be feasible to estimate VAMs with graduation outcomes, enabling us to examine whether schools' impacts on graduation outcomes are closely associated with their impacts on the intermediate holding power outcomes. With these caveats in mind, the Phase 1 evidence indicates that the holding power estimates provide additional information about high schools' impacts on students beyond the information contained in effectiveness estimates for 11th-grade PSSA scores.[42]

Figure V.2. Distribution of School Effectiveness Estimates for 9th- Grade Holding Power



Source: Mathematica calculations based on Pennsylvania data.

Note: Findings are based on a three-cohort model with statewide samples of schools, principals, and students. The sample of principals consists of those who served as principals in every year from 2008-2009 to 2010-2011.

On the other hand, estimates of schools' contributions to attendance outcomes are largely uninformative for identifying effective and ineffective schools, at least in the Phase 1 sample.

[41] In theory, it would have been possible to estimate a VAM for graduation outcomes at the end of spring 2011 with controls for 8th-grade PSSA scores from spring 2007. However, we would not have had data to check the accuracy of the graduation measure—in particular, to check whether students deemed to have graduated actually continued to be enrolled in the following year. Moreover, measuring graduation status at the end of spring 2011 would ignore cases of graduation that occurred as a result of summer school (in summer 2011).

[42] In fact, there is very little correlation—and, in some cases, a slight negative correlation—between the school effectiveness estimates for holding power and those for 11th-grade PSSA scores. For example, correlations between schools' impacts on 11th-grade math PSSA scores and their impacts on holding power range from -0.05 to -0.03, depending on the grade at which holding power is measured. For reading, correlations with the holding power estimates range from -0.05 to 0.

Students enrolled in the 85th percentile school are only one percentage point more likely to be in attendance than they would have been if enrolled in the 15th percentile school. Thus, there is little variation across schools in their effects on attendance rates and, as a result, relatively few (14 percent of) schools are statistically distinguishable from the average on this measure of effectiveness. These findings indicate that the school VAMs for attendance outcomes could not make meaningful, reliable distinctions among schools in the Phase 1 sample. In subsequent phases of the pilot, we will determine whether these findings continue to be observed in a larger pilot sample. In moving to larger samples, the viability of attendance as an outcome will also depend on whether it is measured consistently across districts in the state. Thus, the extent of variation in schools' impacts on attendance and the uniformity with which attendance is measured are the two critical factors for determining whether this outcome should be used in the model evaluation system.

VI. LOOKING AHEAD TO SUBSEQUENT PHASES OF THE PENNSYLVANIA TEACHER AND PRINCIPAL EVALUATION PILOT

As Pennsylvania continues its efforts to improve teacher and principal evaluations in coming years, lessons learned from Phase 1 will be important for the future development of a statewide evaluation system. In the track of the Phase 1 pilot that pertained to measuring effectiveness through the use of student data, Mathematica sought to address research questions related to how value-added models (VAMs) can be used to characterize the effectiveness of teachers and principals at raising student achievement according to multiple outcome measures and whether specific teacher practices relate to larger contributions to student achievement. Consistent with the initial phase of a pilot experiment, the findings from our analyses point to areas of progress and areas in which further attention should be directed during later pilot phases.

We have documented large variation in the estimated effectiveness of teachers and schools based on contributions to growth in assessment scores. The effectiveness estimates can distinguish between educators at the high and low ends of the distribution (provided that student samples sizes are not very small), and we have identified potential outcomes in which the variation in student performance is sufficient to make attributions of effectiveness to individual educators. However, the Phase 1 pilot's small number of teachers was a significant limiting factor for drawing inferences about teacher practices that relate to larger contributions to achievement growth and, to a lesser extent, to the extension of value-added methods beyond grades 4 through 8 that are covered by the Pennsylvania System of School Assessment (PSSA). We also concluded that it is not possible to isolate a principal's contribution to achievement growth using estimates of school-wide effectiveness (though this fact does not necessarily preclude the use of school-level VAM estimates for principal accountability purposes).

Despite these limitations in the Phase 1 analyses, the findings offer much new information and perspective on how to focus the pilot moving forward. Improving the measurement of teacher and principal effectiveness is a difficult challenge for which there is no quick fix, but a challenge that Pennsylvania's leaders have taken on because the ultimate issue at stake is improving student achievement. To conclude our final report from Phase 1 of the Pennsylvania Teacher and Principal Evaluation Pilot, we offer the following list of recommendations in preparation for Phase 2 and later phases that relate broadly to issues of sampling and measurement:

Sample of Teachers and Principals to Be Included in Later Phases

- **Oversample math and reading teachers in grades 4 through 8 and science teachers in grades 4 and 8.** Because a future statewide evaluation model will almost certainly include the PSSA in some capacity, we recommend including a substantial number of teachers in the pilot for whom value-added can be calculated relative to statewide performance. In conjunction with an overall larger Phase 2 sample, we expect that including more of these teachers will be beneficial for identifying relationships between value-added and observation scores.
- **Sample teachers in other grades based on whether outcome and baseline student scores are available in their districts.** For instance, recruit 3rd- or 11th-grade teachers in districts that administer a beginning-year 4Sight assessment in those grades. For purposes of the pilot, Pennsylvania might also seek to incorporate teachers in non-PSSA grades by recruiting among districts that use the same additional standardized

assessments in both non-PSSA and PSSA grades to assess which assessments are most predictive of student achievement as measured by the PSSA.

- **Recruit teachers for subsequent pilot phases to provide for more variation in the observation measure.** Recruit teachers based on the grades and subjects that they teach and on the availability of outcome and baseline assessment data for their students. This method can best facilitate analyses that support pilot goals about understanding relationships between observation scores and value-added. Recruiting only those teachers who are thought to be effective for the pilot might narrow the distribution of observation scores at the end of the year and, consequently, inhibit the pilot's ability to differentiate between the practices of more and less effective teachers.
- **Oversample middle school principals when a new principal evaluation instrument is developed.** Middle school grades are all tested by the PSSA, unlike grade ranges in elementary and high schools. A sample of middle school principals can thus provide for the cleanest analysis of which principal practices are related to larger growth in student achievement because value-added and rubric-based scores will cover the same grades.

Measurement Issues for Later Phases

- **Assess interobserver agreement and observer drift in the observation data.** Later pilot phases should evaluate the interobserver agreement of the teacher observation data. The pilot should also examine the potential for observer drift in evaluation ratings.
- **Evaluate the quality of data linkages in Pennsylvania's student data.** The validity of effectiveness measures based on student data relies fundamentally on the quality of the underlying data system. Most important would be to assess the quality of the student–teacher–principal links. One way to do this would be to provide Phase 2 teachers with the opportunity to verify the data to their classes. It would also be important to assess how educators such as special education teachers are included in the statewide data system and whether a consistent approach is used across districts.
- **Continue to develop nontest outcomes for school value-added measures.** An additional year of data will enable us to extend the concept of holding power to dropout prevention using data on 12th-grade outcomes. Value-added measures for student attendance and core-course completion could be more viable with an expanded sample of Phase 2 schools—or even better, a statewide sample—provided that data are available and that the information is collected consistently across districts.
- **Continue toward defining the evaluation system's structure.** By the conclusion of the pilot, state leaders will put forward a final evaluation model. This model should define the types of teacher and principal effectiveness measures that will be included and how effectiveness information will be integrated across measures. It should also establish standards for accepting elective effectiveness measures proposed by individual school districts. Continued progress toward defining the overall structure will help focus policy goals for the pilot during Phase 2.

REFERENCES

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

Bollinger, C. R., and J. Minier. "On the Robustness of Coefficient Estimates to the Inclusion of Proxy Variables." Working paper. Lexington, KY: University of Kentucky, 2009. http://gatton.uky.edu/faculty/minier/bollingerminier.pdf.

Branch, G., E. Hanushek, and S. Rivkin. "Estimating Principal Effectiveness." Working paper. Dallas, TX: University of Texas at Dallas, 2011.

Coelli, M., and D. Green. "Leadership Effects: School Principals and Student Outcomes." *Economics of Education Review*, forthcoming.

Dhuey, E., and J. Smith. "How Important Are School Principals in the Production of Student Achievement?" Working paper. Toronto, ON: University of Toronto, 2011. Available at http://homes.chass.utoronto.ca/~edhuey/datastore/files/docs/dhuey_smith_princ_nov2011.pdf.

Glazerman, S., S. Loeb, D. Goldhaber, D. Staiger, S. Raudenbush, and G. Whitehurst. "Evaluating Teachers: The Important Role of Value-Added." Washington, DC: Brown Center on Education Policy at Brookings, 2010.

Goldhaber, D., and D. Chaplin. "Assessing the 'Rothstein Falsification Test': Does It Really Show Teacher Value-Added Models Are Biased?" CEDR working paper 2011-5. Seattle, WA: University of Washington, 2011.

Goldhaber, D., and M. Hanson. "Using Performance on the Job to Inform Teacher Tenure Decisions." *American Economic Review: Papers & Proceedings,* vol. 100, no. 2, 2010, pp. 250–255.

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

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

Jacob, B. A., and L. Lefgren. "Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education." *Journal of Labor Economics,* vol. 25, no. 1, 2008, pp. 101–136.

Kane, T. J., J. E. Rockoff, and D. O. Staiger. "What Does Certification Tell Us About Teacher Effectiveness? Evidence from New York City." *Economics of Education Review,* vol. 27, no. 6, 2008, pp. 615–631.

Kane, T. J., and D. Staiger. "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation." NBER Working Paper No. 14607. Cambridge, MA: National Bureau of Economic Research, 2008.

Kane, T. J., and D. O. Staiger. "Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains." Seattle, WA: Bill & Melinda Gates Foundation, Measures of Effective Teaching Project, 2012.

Koedel, C., and J. Betts. "Does Student Sorting Invalidate Value-Added Models of Teacher Effectiveness? An Extended Analysis of the Rothstein Critique." *Education Finance and Policy,* vol. 6, no. 1, 2011, pp. 18–42.

Lane, S., and C. Horner. "Pennsylvania Teacher and Principal Evaluation Pilot Report. Final Report Submitted to the Team Pennsylvania Foundation." Pittsburgh, PA: University of Pittsburgh, 2011.

Lipscomb, S., B. Gill, K. Booker, and M. Johnson. "Estimating Teacher and School Effectiveness in Pittsburgh: Value-Added Modeling and Results, Second Edition." Draft report submitted to the Pittsburgh Public Schools Office of Research, Assessment, and Accountability. Cambridge, MA: Mathematica Policy Research, 2010a.

Lipscomb, S., B. Teh, B. Gill, H. Chiang, and A. Owens. "Teacher and Principal Value-Added: Research Findings and Implementation Practices." Final report submitted to the Team Pennsylvania Foundation. Cambridge, MA: Mathematica Policy Research, 2010b.

Milanowski, A. T., S. M. Kimball, and B. White. "The Relationship Between Standards-Based Teacher Evaluation Scores and Student Achievement: Replication and Extensions at Three Sites." Working paper. Madison, WI: Consortium for Policy Research in Education, 2004.

Pennsylvania Clearinghouse for Education Research. "Teacher Effectiveness: The National Picture and Pennsylvania Context." Philadelphia: Research for Action, 2011.

Rockoff, J. E., and C. Speroni. "Subjective and Objective Evaluations of Teacher Effectiveness." *American Economic Review: Papers & Proceedings,* vol. 100, no. 2, 2010, pp. 261–266.

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

Sartain, L., S. R. Stoelinga, and E. R. Brown. "Rethinking Teacher Evaluation in Chicago." Chicago: Consortium on Chicago School Research, 2011.

Schochet, P., and H. Chiang. "Error Rates in Measuring Teacher and School Performance Based on Student Test Score Gains." Report submitted to the National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences. Princeton, NJ: Mathematica Policy Research, 2010.

Team Pennsylvania Foundation. "Teacher Evaluations Highlighted in Governor's Education Agenda." Published online October 11, 2011. Harrisburg, PA: Team Pennsylvania Foundation, 2011. Available at http://teampa.com.

Tyler, J. H., E. S. Taylor, T. J. Kane, and A. L. Wooten. "Using Student Performance Data to Identify Effective Classroom Practices." *American Economic Review: Papers & Proceedings,* vol. 100, no. 2, 2010, pp. 256–260.

APPENDIX A

TECHNICAL SPECIFICATIONS OF THE VAMS

THIS PAGE LEFT BLANK FOR DOUBLE- SIDED PRINTING

TECHNICAL SPECIFICATIONS OF THE VAMS

In this appendix, we provide a technical description of the value-added models (VAMs). In Section A, we describe the empirical specification for the teacher VAMs and discuss strategies that we employ for enhancing the validity and reliability of results. In Section B, we describe the principal transitions model and its key limitations for usability, and then we describe the school VAM.

A. VAMs for Teachers

The following statistical equation describes the primary teacher VAMs:

$$(1)\quad A_{i,j,g,y} = A_{i,J,g-1,y-1}\beta_1 + \beta_2 A_{i,j,g-2,y-2} + X_{i,y}\gamma + C_{i,j,g,y}\pi + T_{i,j,g,y}\delta + Y_y + e_{i,j,g,y}.$$

In the model, $A_{i,j,g,y}$ is an assessment score for student i in subject j (that is, math, reading, or science) in grade g in year y. For example, the elements of $A_{i,j,g,y}$ could be Pennsylvania System of School Assessment (PSSA) math scaled scores for 5th-grade students across the state or Dynamic Indicators of Basic Early Literacy Skills (DIBELS) scores for 2nd-grade students in Allentown and Cornell. The teacher VAMs use only assessments as outcome measures because available nonassessment measures such as student attendance are indicators of school-wide performance.

$A_{i,J,g-1,y-1}$ is a vector of baseline scores for student i from the prior grade.[43] We use all available prior-grade scores, which vary in number by grade. For example, 5th graders have three scores from 4th grade (that is, PSSA scores in math, reading, and science); however, 7th graders have only two scores from 6th grade (that is, PSSA scores in math and reading). We use prior-grade scores for grade repeaters as well, except that their prior-grade scores come from two years previously. For the outcomes in which a statewide comparison is not possible, $A_{i,J,g-1,y-1}$ can be modified to $A_{i,J,gf,y}$ in Equation (1) if a fall baseline score from grade g is used instead of an end-of-year score from the prior grade. Baseline scores are measured with error. This can bias their coefficient estimates but need not create substantial bias for other coefficient estimates, especially when multiple baseline scores are used as control variables (Bollinger 2009).

When feasible, we also control for a student's same-subject score from two prior grades ago (indicated by $A_{i,j,g-2,y-2}$). Such scores cannot be included in 4th-grade models because state testing begins in 3rd grade. When subjects are not assessed in consecutive grades, as in science and writing, we use a different subject score for $A_{i,j,g-2,y-2}$ instead. For example, the 8th-grade science VAM controls for 6th-grade math and the 8th-grade writing VAM controls for 6th-grade reading. Finally, in VAMs for high school teachers, we use all available 8th-grade PSSA scores as $A_{i,j,g-2,y-2}$ because of the unavailability of assessment data in grades 9 and 10. In all cases, we include linear and quadratic functions of the baseline scores in $A_{i,J,g-1,y-1}$ and $A_{i,j,g-2,y-2}$ to allow for nonlinear relationships between current and prior achievement. This can help address potential issues related to test ceiling effects.

$X_{i,y}$ is a set of control variables for observable student characteristics, $C_{i,j,g,y}$ is a set of classroom-level characteristics, Y_y is a set of year indicator variables, $e_{i,j,g,y}$ is the error term, and $T_{i,j,g,y}$ is a set of

[43] Our VAMs let the data determine the persistence of a teacher's (or a school's) effect on the performance of students in a subsequent year. If the model included only one baseline score control, the coefficient estimate on the baseline score variable would be the degree of empirically determined persistence.

teacher indicator variables. The teacher variables indicate whether a student was taught by a teacher in a given grade, subject, and year, according to Pennsylvania's longitudinal student data. Teachers can be linked to students only if they are listed as the teacher for a course that a student is listed as taking during the school year. We represent the teacher variables as binary 0/1 indicators. Ordinarily, we would use a dosage model that allows for fractional values that sum to one for each student in the event that students are taught by multiple teachers in a subject during the year. Due to the amount of data that enter the statewide teacher VAMs, computational limitations inhibited us from using a dosage model in Phase 1.[44] These constraints are less intensive in the school VAMs—because there are fewer principals than teachers—and, consequently, we use a dosage model there.[45]

The coefficients in β_1, β_2, γ, π, and δ are the estimated relationships between students' assessment scores and each respective variable, controlling for the other factors in the model.[46] There is one δ coefficient for each teacher in the VAM, where each identifies a teacher's contribution to student learning—the extent to which the actual achievement of students tends to be above or below what is expected for the average teacher. We define the average VAM score to be a zero value, but this does not mean that student learning is zero for the teacher with the average VAM score. Rather, it means that a positive VAM score represents above-average teacher performance and a negative VAM score represents below-average teacher performance.

The reference point for determining the average teacher contribution depends on the sample of teachers in the model. If the data include students and teachers across Pennsylvania (for example, as in PSSA math grade 5), the VAM estimates would be calculated relative to the contribution of the average teacher in Pennsylvania in that grade and subject. If the data include only students and teachers from a particular grade in pilot districts (for example, 2nd-grade DIBELS in Allentown and Cornell), the VAM scores would be calculated relative to the contribution of the average teacher in that grade and subject at the particular pilot district(s). Our VAMs include both statewide and Phase 1 district samples, with the primary determining factor being the level at which the outcome and baseline data are available. Appendix B contains detailed information on data elements and samples.

In estimating the VAMs, we take the following additional steps to enhance the accuracy and reliability of the results:

- **Convert assessment scores into standard scores.** VAM estimates reported in assessment units (for example, PSSA scaled score points) are not comparable across assessments, grades, subjects, or years. Before estimating a VAM, we map assessment scores to a standard measure, called a z-score, by subtracting the average annual score from individual scores (by grade and school year) and then dividing by the standard

[44] For example, the statewide teacher VAM for 5th-grade math takes 14 hours to run on a Windows 7 personal computer using Stata 11 2-core Multi-Processor (MP) Edition when storing the teacher variables as binary indicators. The memory needed to estimate the VAM would quadruple if teacher variables were allowed to take fractional values.

[45] In the school models, the dosage is split evenly across principals leading all schools a student attended during the year. Because Pennsylvania's longitudinal data system did not include attendance data for us to use in this report, we are not able to account for any time that students are enrolled outside Pennsylvania public schools during the academic year. The approximate run time for these models is one hour using Windows 7 and Stata 11 2-core MP Edition.

[46] The standard errors adjust for clustering of observations by student.

deviation of scores. Expressing scores in this way enables us to interpret above-average scores in terms of how close to the average most students tended to fall. Appendix C reports VAM results in z-score terms. Estimates are converted back to score points for the reporting of results in Chapters III and V.

- **Adjust estimates based on their precision.** Consistent with the research literature, the VAMs use a procedure known as empirical Bayes estimation or shrinkage to address the fact that, among teachers with the same level of true performance, those with fewer students in the estimation sample face a greater likelihood that their students happen, by chance, to have atypically high or low learning growth driven by other factors. Teachers with fewer students—that is, those with less precise estimates—will tend to be over-represented at both the high and low ends of the estimated performance distribution for reasons other than their effectiveness. Shrinkage adjustments account for the fact that estimates with greater precision carry greater strength of information about teachers' true performance levels. The adjusted estimate is a weighted average of the individual initial estimate and the mean estimate across teachers, with more precise initial estimates receiving greater weight. In essence, teachers are assumed to be average in performance until evidence justifies a different conclusion. To further minimize the risk of making erroneous conclusions on the basis of imprecise estimates, we limit analyses to teachers who taught more than 10 students during the year.

- **Incorporate observations on students across multiple years.** Our primary teacher VAMs include students taught by a teacher in the past three years—that is, the number of current and prior student cohorts who contribute to the estimate—whenever three years of data are available. Multicohort VAM estimates are less prone to random fluctuations that stem from a teacher being assigned a few students with unusually high or low learning growth. The VAM estimates can therefore detect performance differences with greater reliability. Researchers have also found that multicohort VAMs are less prone to systematic fluctuations in scores as well, meaning that they might have greater validity, too. Multicohort VAMs also better distinguish teacher effects from the effects of students' peers in the classroom, which is impossible to determine in a single-cohort model unless teachers teach in multiple classrooms during the year. We report estimates from multicohort VAMs as follows. First, we estimate the VAM based on all students and teachers in the included school years. We then restrict the resulting set of teacher estimates to those for teachers with students in the outcome subject and grade in the latest year covered by the VAM and with more than 10 students overall during the sample period. We next apply the empirical Bayes calculation to this subset of estimates and then center the resulting effectiveness measures on a zero value. We report estimates for teachers with students from all three prior years, rather than for all teachers, including those with fewer than three years of teaching data. In this report, we also provide estimates from several one-cohort VAMs, for the purpose of comparing the results.

- **Adjust some teacher effectiveness estimates for district and school factors.** In a diagnostic analysis, we adjust the teacher effectiveness estimates for district and school factors by subtracting the mean teacher effect in each district or school from each individual estimate. This adjustment—which has the same effect as adding district or school fixed effects to the VAM itself—has the potential to provide better controls for district- or school-level influences on teacher performance that are external to teachers. We do not make this adjustment to our primary teacher VAMs, however, because it

means that teachers are compared only with other teachers in their same district or school. It might also under-represent true differences in teacher effectiveness to the extent that highly effective (or ineffective) teachers tend to cluster together.

B. VAMs for Principals

1. Principal Transitions Model

The principal transitions model is based on leadership transitions between principals within the same school (Branch et al. 2011; Dhuey and Smith 2011; Coelli and Green, forthcoming). Thus, this model is fundamentally based on comparing principals who lead the same school (at different times) during the analysis period. As long as the mix of existing abilities within a school's teaching staff remains constant during the analysis period, comparing principals who have led the same school will effectively remove the influence of existing teacher abilities from the estimate of a principal's value-added.[47] A related advantage of this method is that it also removes the influence of other school-specific factors that remain constant during the analysis period, such as neighborhood quality.

The principal transitions model can be represented by the following equation:

$$(2)\quad A_{i,j,g,y} = A_{i,J,g-1,y-1}\beta_1 + \beta_2 A_{i,j,g-2,y-2} + X_{i,y}\gamma + P_{i,j,g,y}\varphi + S_{i,j,g,y}\delta + Y_y + e_{i,j,g,y},$$

where $P_{i,j,g,y}$ is a set of principal variables, $S_{i,j,g,y}$ is a set of school variables, and all other variables are defined as in Equation (1). The coefficients, φ, on the principal variables are the estimates of principal effectiveness. Importantly, the school variables control for any school-specific influences on student achievement—potentially including existing teacher abilities—that do not change during the analysis period.[48] As discussed in Chapter V, the principal transitions model is likely to have a high degree of internal validity, but estimates from this model can be compared only among principals who have led schools connected by leadership transitions during the analysis period. For this reason, we do not consider this model to be a viable method for generating effectiveness measures in actual principal evaluations.

We use the principal transitions model as a point of comparison to an alternative model that bundles together the combined value-added of educators at the school. This latter model, which we call the school VAM, does not control for the mix of existing teacher abilities or other school factors outside a principal's control but can be applied to all principals in Pennsylvania.

[47] Even if the composition of the teaching staff changes concurrently with the leadership transition, the existing abilities of teachers remaining at the school will be removed from the principal effectiveness estimates.

[48] An alternative way to control for existing teacher abilities is to include teacher variables, rather than school variables, in Equation (2). This approach estimates principals' effects by examining how teacher value-added changes when a teacher transitions between principals. We chose not to use this approach because teachers' transitions between principals often occur as a result of teacher transfers between schools; a change in the value added of transferring teachers could be due to differences in school-specific factors other than just the change in principal.

2. School Model

The school VAM that we estimate is conceptually similar to the teacher VAM discussed previously and has the same basic empirical specification:

$$(3)\ A_{i,j,g,y} = A_{i,J,g-1,y-1}\beta_1 + \beta_2 A_{i,j,g-2,y-2} + X_{i,y}\gamma + P_{i,j,g,y}\delta + Y_y + e_{i,j,g,y}.$$

Most of the variables are defined in the same way as in Equation (1). For the school VAM, $P_{i,j,g,y}$ now denotes a set of principal variables, and hence the δ terms identify total contributions to student learning by educators at a principal's school(s), including the principal's own contribution. Most school VAMs control for schools rather than for principals. We chose to use principal variables because the pilot's focus is on principal evaluation; thus, the estimates measure the school value-added at a principal's school or schools. A conceptual distinction between school and teacher VAMs is that the baseline score in a teacher VAM is measured before each student entered the teacher's class, whereas the baseline score in a school VAM typically comes from the preceding year regardless of whether the student attended a school led by the same or a different principal in that year.

In applying school VAMs, we also use several nonassessment outcomes, such as attendance rates and rates of staying in school (see Appendix B). The VAMs are generally the same for nonassessment outcomes as for assessment outcomes.[49] However, one difference is that the baseline achievement measures might capture a somewhat different aspect of achievement than the outcome measure. For instance, when evaluating high schools' impacts on students' rates of staying in school, we control for students' assessment scores from 8th grade. Although the elements of student performance captured by these two variables are not identical, 8th-grade test scores might nevertheless be strongly predictive of staying in school, and controlling for these scores will reduce the likelihood that comparisons across schools could be biased by differences in their students' incoming risk of dropping out.

We apply all of the aforementioned steps for enhancing the validity and reliability of VAMs, including standardization of assessment scores, use of dosage to apportion responsibility for the outcomes of mobile students, adjusting VAM estimates for their level of precision, and exploring the implications of using multiple years of student growth data. For the school VAM approach, if we incorporate multiple years of student growth data, then there are principals who work in multiple schools during the analysis period. The principals' final VAM scores will be averaged over all of the relevant school–principal combinations.[50]

Despite the similarities between school and teacher VAMs, there are two main substantive differences between these models. The first was mentioned earlier: the baseline scores used for most grades in the school models are not pretreatment measures—that is, measures from a period before students' enrollment in a specified school. Except in entry grades for middle and high schools,

[49] Although most VAMs in practice have focused on assessment outcomes, the methodology is very similar for nonassessment outcomes, so we also apply the terminology of *value-added models* to these latter methods as well.

[50] Several schools are led by more than one principal at the same time. If a group of principals is observed to have led the same school at the same time in all years of the VAM, then they are assigned a single effectiveness estimate.

students are generally served by the same principal both in the current grade (that is, the grade to which a set of VAM estimates apply) and in the prior grade, when baseline scores are measured. In contrast, baseline scores in the teacher models are typically pretreatment, given that teachers generally do not teach the same students in multiple grades. In prior analysis for the Pittsburgh Public Schools, we found that school VAM estimates were similar regardless of whether prior-grade baseline scores or pre-entry baseline scores were used (Lipscomb et al. 2010a). Our models therefore use prior-grade baseline controls because this approach can be applied consistently across grades and schools and allows the inclusion of a larger number of students.

A related limitation of the school VAM applies specifically to elementary schools. Because PSSA assessments begin in grade 3, VAMs relying solely on PSSA scores for both baseline and outcome measures provide no information about the value-added produced from kindergarten entry through the end of 3rd grade. In other words, unless school districts administer additional assessments in the early elementary grades (as some do), four years of schooling are invisible to VAMs that rely on tests beginning in grade 3.

APPENDIX B

DATA SOURCES AND SAMPLE CHARACTERISTICS

THIS PAGE LEFT BLANK FOR DOUBLE- SIDED PRINTING

DATA SOURCES AND SAMPLE CHARACTERISTICS

In this appendix, we describe the data that are used in the report. In Section A, we list the source for each data element. In Section B, we compare the characteristics of students in the pilot districts to those of students attending other school districts. In Section C, we provide the baseline and analysis sample sizes for students, teachers, and principals.

A. Data Sources

Nearly all data for this report come from the statewide Pennsylvania Department of Education (PDE) data and from the four Phase 1 pilot districts. In Table B.1, we summarize data elements by source. PDE's Pennsylvania Information Management System (PIMS) is the source for student characteristics and most information on linkages between students and their teachers, principals, and schools. The exception is for student–teacher links for students attending Pittsburgh Public Schools, which are largely missing in PIMS. For Pittsburgh, we used the district's own records instead.[51] State assessment data come from PDE's Bureau of Assessment and Accountability (BAA) and PIMS. The remaining variables were obtained from pilot districts directly.

[51] The missing data pertain to the student–course records (template 490). We identified the missing data problem for Pittsburgh through our involvement in Pittsburgh's Empowering Effective Teachers project and were able to use the district's own records because Mathematica already had access to them.

Table B.1. Data Sources

Agency	Data Element	School Years
Pennsylvania (PIMS)	Student background (template 0320)	2007-2008 to 2010-2011
	Student-course links (template 0490)	2008-2009 to 2010-2011
	Teacher-course links (template 0410)	2008-2009 to 2010-2011
	Principal-school links (template 0630)	2008-2009 to 2010-2011
	Course description (template 0310)	2008-2009 to 2010-2011
	PSSA scaled scores (all subjects)	2007-2008 to 2009-2010
Pennsylvania (BAA)	PSSA scaled scores (all subjects)	2010-2011
	PSSA-M scaled scores (all subjects)	2009-2010 to 2010-2011
Allentown	Student attendance	2009-2010 to 2010-2011
	Core courses attempted and passed	2009-2010 to 2010-2011
	4Sight scores	May 2010, Sept. 2010
	DIBELS scores	May 2010, Sept. 2010, May 2011
	Progress assessment scores	May 2010, Oct. 2010, May 2011
	Teacher observation rubric scores	Spring 2011
Cornell	Core courses attempted and passed	2009-2010 to 2010-2011
	4Sight scores	May 2010, Sept. 2010
	DIBELS scores	April 2010, Sept. 2010, April 2011
	Teacher observation rubric scores	Spring 2011
Mohawk	Student attendance	2009-2010 to 2010-2011
	Core courses attempted and passed	2009-2010 to 2010-2011
	4Sight scores	May 2010, Sept. 2010
	Teacher observation rubric scores	Spring 2011
Northwest Tri-County	Student attendance	2009-2010 to 2010-2011
	Teacher observation rubric scores	Spring 2011

BAA = Bureau of Assessment and Accountability; DIBELS = Dynamic Indicators of Basic Early Literacy Skills; PIMS = Pennsylvania Information Management System; PSSA = Pennsylvania System of School Assessment; PSSA-M = PSSA-Math.

B. Descriptive Statistics

In Table B.2, we report the baseline sample means for several student characteristics that are used in the VAMs. These data come from the PIMS for 2010–2011. The first two columns show averages across all students in nonpilot and pilot districts. The remaining four columns apply to individual pilot districts.

Table B.2. Descriptive Statistics on Student Characteristics, 2010- 2011

Variable	Nonpilot PA Average	Pilot District Average	Pilot Districts			
			Allentown	Cornell	Mohawk	Northwest
Math PSSA, Grade 5 (scaled score)	1,468.5	**1,402.8**	**1,395.3**	**1,420.6**	**1,520.1**	*
Math PSSA, Grade 11 (scaled score)	1,373.8	**1,210.6**	**1,193.9**	**1,261.5**	1,372.7	**1,110.5**
Reading PSSA, Grade 8 (scaled score)	1,505.7	**1,322.2**	**1,289.3**	**1,451.9**	**1,470.2**	**1163.2**
Science PSSA, Grade 4 (scaled score)	1,452.9	**1,329.6**	**1,316.4**	**1,375.7**	**1,506.0**	*
Female (%)	48.4	**47.0**	**47.5**	50.3	47.7	**31.0**
White (%)	71.3	**26.2**	**16.4**	70.7	**97.3**	68.4
African American (%)	15.8	16.2	**17.3**	14.2	*	**24.3**
Hispanic (%)	8.0	**54.7**	**63.6**	**2.1**	*	**4.6**
Asian and Pacific Islander (%)	3.2	**1.4**	**1.5**	*	**0.6**	*
Multiracial or Other Race/Ethnicity (%)	1.7	1.6	**1.1**	**13.1**	1.2	1.9
Free Lunch Eligibility (%)	34.0	**70.3**	**75.3**	**52.4**	**28.9**	**54.6**
Reduced-Price Lunch Eligibility (%)	5.6	**8.9**	**9.3**	**9.6**	5.9	5.2
English-Language Learner (%)	2.8	**10.1**	**11.7**	*	*	**1.4**
Special Education (%)	16.0	**15.5**	**13.0**	17.2	**11.4**	**78.0**
Grade Repeater (%)	2.6	**5.1**	**5.2**	1.9	**1.0**	**14.1**
Number of Students (1,000's)	1,617.6	22.1[a]	18.9	0.7	1.6	0.9

Source: Mathematica calculations based on Pennsylvania student data.

Note: **Bold** indicates a statistically significant mean difference between the nonpilot district average and the pilot district average at the 5 percent level. Descriptive statistics may differ for analysis samples based on the characteristics of the students included in each model. Pilot districts account for 1.3 percent of all students in the state.

* Indicates that a sample mean is withheld because it includes 10 or fewer students.

[a] The number 22.1refers to the sum of students across the four pilot districts (in thousands).

As indicated in the table, the sample characteristics of students in the four pilot districts differ from the characteristics of students in other Pennsylvania districts in terms of most of the observable demographic variables. That is, the Phase 1 districts are not representative of the state in terms of their student populations. The pilot district averages (column 2) are primarily influenced by Allentown, given its size compared with the other pilot districts, but each pilot district has important differences with nonpilot districts in terms of the characteristics of its students. Given these differences in baseline characteristics, we recommend interpreting findings based on analyses of pilot data as suggestive for future work involving larger Pennsylvania samples but not as representative of Pennsylvania students, teachers, or schools.

C. Baseline and Analysis Sample Sizes

In Table B.3, we describe the baseline and analysis samples for students in the teacher and school VAMs. By baseline sample, we mean the number of students that have a nonmissing value of the outcome variable for a particular VAM. The analysis sample includes the subset of those students with nonmissing data on prior scores, student characteristics, and teacher–principal links.

Table B.3. Baseline and Analysis Student Sample Sizes for Teacher and School VAMs, by Outcome

Outcome	Description of the Sample	Number of Cohorts in the Main Model	Number of Students with Nonmissing Values of the Outcome Measure	Number of Students in the Analysis Sample for Teacher VAMs	Number of Students in the Analysis Sample for School VAMs
Math PSSA, Grade 3	A,C,M	1	1,676	1,291	1,309
Math PSSA, Grade 4	PA	3	361,916	304,013	337,017
Math PSSA, Grade 5	PA	3	364,180	287,889	320,734
Math PSSA, Grade 6	PA	3	364,192	295,247	317,807
Math PSSA, Grade 7	PA	3	367,897	311,298	322,251
Math PSSA, Grade 8	PA	3	373,308	311,991	326,470
Math PSSA, Grade 11	A,C,M (T)	1	1,450	725	--
Math PSSA, Grade 11	PA (P)	2	242,819	--	196,976
Reading PSSA, Grade 3	A,C,M	1	1,666	1,291	1,310
Reading PSSA, Grade 4	PA	3	361,376	305,736	336,932
Reading PSSA, Grade 5	PA	3	363,678	287,611	319,966
Reading PSSA, Grade 6	PA	3	363,717	300,605	317,023
Reading PSSA, Grade 7	PA	3	367,273	312,985	321,469
Reading PSSA, Grade 8	PA	3	372,676	315,056	325,645
Reading PSSA, Grade 11	A,C,M (T)	1	1,450	730	--
Reading PSSA, Grade 11	PA (P)	2	242,371	--	196,834
Writing PSSA, Grade 5	PA	3	361,100	286,229	318,422
Writing PSSA, Grade 8	PA	3	369,574	313,409	323,880
Writing PSSA, Grade 11	A,C,M (T)	1	1,382	718	--
Writing PSSA, Grade 11	PA (P)	2	237,615	--	194,537
Science PSSA, Grade 4	PA	3	360,596	290,214	336,015
Science PSSA, Grade 8	PA	3	370,052	307,544	324,493
Science PSSA, Grade 11	A,C,M (T)	1	1,359	693	--
Science PSSA, Grade 11	PA (P)	2	237,189	--	193,955
Math Progress Assess., Grade 2	A	1	1,176	865	870
Writing Progress Assess., Grade 1	A	1	1,225	899	903
Writing Progress Assess., Grade 2	A	1	1,212	894	900
Writing Progress Assess., Grade 3	A	1	1,126	870	874
DIBELS (NWF), Grade 1	A	1	1,434	968	987
DIBELS (PSF), Grade 1	A	1	1,434	968	987
DIBELS (ORF), Grade 2	A,C	1	1,055	831	836
Attendance Rate, Grades 4-12	A,M,N (P)	1	11,787	--	9,339
Holding Power, Grade 9	PA (P)	3	294,343	--	244,709
Holding Power, Grade 10	PA (P)	2	292,921	--	229,602
Holding Power, Grade 11	PA (P)	1	136,785	--	107,972

Source: Mathematica calculations based on Pennsylvania student data and student data from pilot districts.

Note: Sample sizes refer to student-school year observations. Students are counted more than once if they appear in a sample in multiple years. The analysis sample for an outcome measure is the sample that is used for estimating a VAM.

A = Allentown; C = Cornell; M = Mohawk; N = Northwest; P = principal; PA = Pennsylvania; T = teacher.

-- indicates that sample size information is not available because a model was not estimated.

As described in Appendix A, the samples differ by outcome because some VAMs can include students from across the Commonwealth, whereas others are limited to pilot districts. The sample sizes also differ based on the number of student cohorts (up to three) that can be included. On average, analysis samples for school VAMs are 11 percent smaller than baseline samples in grades 4 to 8. Analysis samples for teacher VAMs in grades 4 to 8 are 17 percent smaller than baseline samples, on average. The primary source of sample reduction in the school VAMs is students who are missing at least one prior test. In contrast, only a small number of students are excluded for

other reasons. Sample exclusions in the teacher VAMs reflect two primary factors: students with missing score history (that is, the same reason as in the school VAMs) and missing teacher links. As described earlier in this appendix, we have partially addressed this source of sample loss by using Pittsburgh's own records on student–teacher links. However, as indicated by the difference in sample sizes between the last two columns in Table B.3, Pennsylvania can increase the number of students and teachers in the VAMs for future years by improving the quality of student–teacher links.

In Table B.4, we report the number of teachers and principals with VAM estimates by outcome and by whether a VAM includes a single cohort or multiple cohorts of students. Although multicohort VAMs include more students, fewer estimates are reported because we output results only for teachers or principals linked to students during the entire multiyear period.[52] For example, in the three-cohort VAM for 5th-grade math based on the PSSA, we report estimates for teachers with 5th-grade math students in 2008–2009, 2009–2010, and 2010–2011. We do not report estimates from that VAM for other teachers, such as new 5th-grade math teachers in 2010–2011. In future years if desired, it would be possible to report all estimates from a multicohort VAM regardless of the number of cohorts of students it includes for a given teacher.

[52] As described in Appendix A, we also report estimates only for teachers and principals who can be linked with more than 10 students in the analysis because estimates based on very small numbers of students are likely to have low precision. This requirement applies for both single-cohort and multicohort VAMs.

Table B.4. Number of Teachers and Principals with VAM Estimates Reported from Multicohort and Single- Cohort VAMs

Outcome	Description of the Sample	Number of Cohorts in the Main Model	Teachers with VAM Estimates		Principals with VAM Estimates	
			Single Cohort	Multi-Cohort	Single Cohort	Multi-Cohort
Math PSSA, Grade 3	A,C,M	1	69	--	16	--
Math PSSA, Grade 4	PA	3	--	3,075		1,167
Math PSSA, Grade 5	PA	3	4,103	2,836	1,336	1,079
Math PSSA, Grade 6	PA	3	--	1,994	--	758
Math PSSA, Grade 7	PA	3	--	1,403	--	581
Math PSSA, Grade 8	PA	3	1,685	1,471	--	580
Math PSSA, Grade 11	A,C,M (T)	1	29	--	--	--
Math PSSA, Grade 11	PA (P)	2	--	--	626	577
Reading PSSA, Grade 3	A,C,M	1	69	--	16	--
Reading PSSA, Grade 4	PA	3	--	3,126	--	1,167
Reading PSSA, Grade 5	PA	3	4,167	2,907	--	1,079
Reading PSSA, Grade 6	PA	3	--	2,446	--	758
Reading PSSA, Grade 7	PA	3	--	1,749	--	581
Reading PSSA, Grade 8	PA	3	1,916	1,717	755	580
Reading PSSA, Grade 11	A,C,M (T)	1	19	--	--	--
Reading PSSA, Grade 11	PA (P)	2	--	--	--	578
Writing PSSA, Grade 5	PA	3	--	2,908	--	1,077
Writing PSSA, Grade 8	PA	3	--	1,711	--	579
Writing PSSA, Grade 11	A,C,M (T)	1	19	--	--	--
Writing PSSA, Grade 11	PA (P)	2	--	--	--	574
Science PSSA, Grade 4	PA	3	4,187	2,854	1,427	1,166
Science PSSA, Grade 8	PA	3	1,313	1,035	--	581
Science PSSA, Grade 11	A,C,M (T)	1	17	--	--	--
Science PSSA, Grade 11	PA (P)	2	--	--	--	574
Math Progress Assess., Grade 2	A	1	46	--	12	--
Writing Progress Assess., Grade 1	A	1	50	--	13	--
Writing Progress Assess., Grade 2	A	1	47	--	13	--
Writing Progress Assess., Grade 3	A	1	48	--	14	--
DIBELS (NWF), Grade 1	A	1	53	--	13	--
DIBELS (PSF), Grade 1	A	1	53	--	13	--
DIBELS (ORF), Grade 2	A,C	1	44	--	15	--
Attendance Rate, Grades 4-12	A,M,N	1	--	--	29	--
Holding Power, Grade 9	PA	3	--	--	--	612
Holding Power, Grade 10	PA	2	--	--	--	612
Holding Power, Grade 11	PA	1	--	--	690	--

Source: Mathematica calculations based on Pennsylvania student data.

Note: Teachers and principals are included in multiple rows if they have students in multiple grades.

A = Allentown; C = Cornell; M = Mohawk; N = Northwest; P = principal; PA = Pennsylvania; T = teacher.

-- indicates that sample size information is not available because a model was not estimated.

APPENDIX C

TECHNICAL RESULTS FROM VALUE- ADDED ANALYSES

THIS PAGE LEFT BLANK FOR DOUBLE- SIDED PRINTING

TECHNICAL RESULTS FROM VALUE- ADDED ANALYSES

In this appendix, we provide the full technical results from value-added models (VAMs) applied for estimating teacher and principal effectiveness. The tables are sequenced to correspond with the presentation of findings in Chapters III and V.

Table C.1. Sample Characteristics of Outcome Measures and Teacher VAMs Based on State Samples

	Distribution of Student-Level Outcome Variable		Characteristics of Teacher VAMs and Estimates			
Outcome	Mean	Standard Deviation	Adjusted R-Squared	85th Minus 50th Percentile of VAM Estimates (in z-score units)	Mean Standard Error (in z-score units)	Percentage of VAM Estimates that Are Statistically Distinguishable from the Average
Math PSSA, Grade 4	1,475	221	0.68	0.23	0.07	51.6
Math PSSA, Grade 5	1,479	223	0.78	0.20	0.06	52.0
Math PSSA, Grade 6	1,501	237	0.79	0.19	0.05	58.6
Math PSSA, Grade 7	1,502	239	0.81	0.17	0.04	61.7
Math PSSA, Grade 8	1,454	220	0.81	0.16	0.04	59.2
Reading PSSA, Grade 4	1,387	212	0.68	0.16	0.07	38.7
Reading PSSA, Grade 5	1,351	209	0.73	0.16	0.06	38.9
Reading PSSA, Grade 6	1,397	222	0.74	0.11	0.06	32.7
Reading PSSA, Grade 7	1,428	214	0.74	0.11	0.06	32.2
Reading PSSA, Grade 8	1,519	247	0.75	0.09	0.05	30.5
Writing PSSA, Grade 5	1,342	258	0.53	0.30	0.09	57.6
Writing PSSA, Grade 8	1,412	262	0.55	0.21	0.07	48.7
Science PSSA, Grade 4	1,462	182	0.66	0.22	0.07	49.8
Science PSSA, Grade 8	1,324	198	0.73	0.14	0.04	57.2

Source: Mathematica calculations based on Pennsylvania student data.

Notes: Findings on PSSA scores are based on three-cohort models with statewide samples of teachers and students. For three-cohort models, the sample of teachers consists of those who served as teachers in every year from 2008-2009 to 2010-2011. Teachers' VAM estimates are based on students in their classrooms at any time during the specified analysis periods. One z-score unit is equal to one standard deviation of student outcomes.

PSSA = Pennsylvania System of School Assessment; VAM = value-added model.

Table C.2. Estimated Regression Coefficients from Selected Three- Cohort PSSA Teacher VAMs

	Math, Grade 5 (in z-score units)		Reading, Grade 8 (in z-score units)		Science, Grade 4 (in z-score units)	
	Coefficient	t-stat	Coefficient	t-stat	Coefficient	t-stat
Math, Prior Grade	0.4888	254.98	0.1877	114.21	0.2794	160.59
Math, Prior Grade ^2	-0.0070	-5.04	-0.0003	-0.25	-0.0393	-29.16
Reading, Prior Grade	0.0722	41.47	0.3767	195.30	0.4624	252.01
Reading, Prior Grade ^2	0.0311	25.07	-0.0287	-21.25	0.0270	18.74
Science, Prior Grade	0.1263	72.57				
Science, Prior Grade ^2	0.0184	15.00				
Outcome Subject, 2-Prior Grade	0.2244	133.96	0.2718	149.51		
Outcome Subject, 2-Prior Grade ^2	-0.0263	-21.15	-0.0130	-9.93		
Free Meals	-0.0197	-11.63	-0.0216	-11.36	-0.0460	-19.94
Reduced-Price Meals	-0.0069	-7.39	-0.0108	-11.36	-0.0148	-12.86
English-Language Learner	0.0055	2.53	-0.0003	-0.11	-0.0098	-3.10
Specific Learning Disability	-0.0350	-30.74	-0.0221	-17.15	0.0177	12.39
Speech or Language Impairment	-0.0012	-1.40	-0.0020	-2.29	-0.0103	-9.20
Emotional Disturbance	-0.0113	-9.20	-0.0096	-6.51	0.0041	1.79
Intellectual Disability	-0.0023	-1.13	-0.0076	-4.95	0.0003	0.12
Autism	-0.0078	-6.34	-0.0058	-4.66	0.0012	0.52
Physical/Sensory Disability	-0.0036	-1.56	-0.0052	-3.01	-0.0045	-1.84
Other Impairment	-0.0178	-17.89	-0.0115	-11.53	-0.0041	-3.39
Mobility	-0.0069	-7.07	-0.0044	-4.45	-0.0016	-1.37
Grade Repeater	0.0045	3.39	0.0076	5.53	0.0162	12.20
Behind	-0.0006	-0.55	-0.0054	-4.92	-0.0049	-3.68
Age	-0.0247	-24.69	-0.0029	-2.78	-0.0013	-1.03
PSSA-Modified (outcome)	0.0711	77.03	0.0534	59.75		
PSSA-Modified (baseline)	-0.0242	-25.92	-0.0034	-3.44		
Female	-0.0064	-6.83	0.0701	73.12	-0.0698	-62.06
Asian/Pacific Islander	0.0263	25.57	0.0099	9.69	0.0057	4.78
African American	-0.0053	-4.07	0.0079	6.06	-0.0486	-30.12
Hispanic	0.0010	0.82	0.0047	4.01	-0.0229	-15.94
Other Race/Ethnicity	-0.0012	-1.28	0.0009	0.88	-0.0044	-3.74
Class Avg.: Free Meals	-0.0059	-1.74	-0.0414	-15.37	-0.0236	-5.30
Class Avg.: Reduced-Price meals	0.0015	1.05	-0.0089	-7.24	-0.0027	-1.48
Class Avg.: English-Language Learner	0.0047	2.14	-0.0088	-5.11	0.0015	0.51
Class Avg.: Special Education	-0.0004	-0.25	-0.0232	-12.66	0.0016	0.79
Class Avg.: Female	0.0027	2.10	0.0107	9.46	0.0016	1.01
Class Avg.: Asian/Pacific Islander	-0.0015	-0.71	0.0126	7.61	-0.0072	-2.84
Class Avg.: African American	0.0009	0.20	-0.0029	-0.90	-0.0408	-7.52
Class Avg.: Hispanic	-0.0029	-0.77	-0.0027	-1.00	-0.0234	-4.99
Class Avg.: Other Race/Ethnicity	0.0016	1.04	-0.0002	-0.20	-0.0093	-4.59
Class Size	-0.0091	-2.92	0.0097	4.82	-0.0065	-1.56
Class Size x Emotional Disturbance	-0.0018	-2.13	-0.0002	-0.19	-0.0045	-2.11
Class Size x Intellectual Disability	-0.0007	-0.32	-0.0010	-0.65	-0.0045	-1.61
Class Size x Autism	-0.0006	-0.66	0.0018	1.67	-0.0046	-2.04
Class Size x Physical/Sensory	0.0007	0.29	0.0029	1.76	0.0020	0.87
Class Size x Free Meals	-0.0036	-2.23	-0.0060	-3.11	-0.0026	-1.14
Class Size x English-Language Learner	0.0001	0.03	0.0026	0.93	0.0018	0.60

Source: Mathematica calculations based on Pennsylvania student data.

Notes: T-statistics that exceed 1.96 in absolute value are statistically significant at the 5 percent level. All variables are expressed in standard deviation units with a mean of zero. These regressions include indicator variables for each teacher and school year, and no intercept. One z-score unit is equal to one standard deviation of student outcomes.

Table C.3. Effect Sizes for Three- Cohort Teacher VAMs Expressed in Terms of One Year of Learning

	85th Minus 50th Percentile of VAM Estimates in Z-score Units		Average Annual Gain on Nationally Normed Tests in Z-score Units (from Hill et al. 2008)		85th Minus 50th Percentile of VAM Estimates in Terms of One Year of Learning for a Typical Student	
Grade	Math	Reading	Math	Reading	Math	Reading
4	0.23	0.16	0.52	0.36	0.44	0.44
5	0.20	0.16	0.56	0.40	0.36	0.40
6	0.19	0.11	0.41	0.32	0.46	0.34
7	0.17	0.11	0.30	0.23	0.57	0.48
8	0.16	0.09	0.32	0.26	0.50	0.35

Source: Table C.1 and Hill et al. (2008) Table 1.

Note: One z-score unit is equal to one standard deviation of student outcomes. A difference in VAM estimates expressed in terms of one year of learning equals the difference expressed in z-score units divided by the average annual gain in z-score units.

VAM = value-added model.

Table C.4. Sample Characteristics of Outcome Measures and Teacher VAMs Based on Phase 1 Samples

	Distribution of Student-Level Outcome Variable		Characteristics of Teacher VAMs and Estimates			
Outcome	Mean	Standard Deviation	Adjusted R-Squared	85th Minus 50th Percentile of VAM Estimates (in z-score units)	Mean Standard Error (in z-score units)	Percentage of VAM Estimates that Are Statistically Distinguishable from the Average
DIBELS (NWF), Grade 1[a]	39	22	0.42	0.36	0.16	41.5
DIBELS (PSF), Grade 1[a]	40	15	0.68	0.50	0.12	56.6
DIBELS (ORF), Grade 2[a]	85	37	0.81	0.13	0.09	18.2
Writing Progress Assessment, Grade 1[b]	76	14	0.52	0.34	0.15	38.0
Writing Progress Assessment, Grade 2[b]	76	15	0.59	0.35	0.13	48.9
Math Progress Assessment, Grade 2[b]	73	16	0.68	0.29	0.12	34.8
Writing Progress Assessment, Grade 3[b]	75	19	0.56	0.26	0.14	27.1
Math PSSA, Grade 3[c]	1,301	186	0.69	0.30	0.12	43.5
Reading PSSA, Grade 3[c]	1,286	158	0.71	0.23	0.12	18.8
Math PSSA, Grade 11[c]	1,280	211	0.79	0.08	0.09	3.4
Reading PSSA, Grade 11[c]	1,277	235	0.74	0.05	0.08	0.0
Writing PSSA, Grade 11[c]	1,451	238	0.52	0.10	0.12	0.0
Science PSSA, Grade 11[c]	1,189	89	0.71	0.03	0.08	0.0

Source: Mathematica calculations based on Pennsylvania student data and student data from pilot districts.

Note: Findings are based on one-cohort models in which the sample of teachers consists of those who served as teachers in 2010-2011. One z-score unit is equal to one standard deviation of student outcomes.

[a] Findings are based on students and teachers in Allentown and Cornell.
[b] Findings are based on students and teachers in Allentown.
[c] Findings are based on students and teachers in Allentown, Cornell, and Mohawk.

DIBELS = Dynamic Indicators of Basic Early Literacy Skills; NWF = nonsense word frequency; ORF = oral reading fluency; PSF = phoneme segmentation fluency; PSSA = Pennsylvania System of School Assessment; VAM = value-added model.

Table C.5. Sample Characteristics of Outcome Measures and School VAMs Based on State Samples

	Distribution of Student-Level Outcome Variable		Characteristics of School VAMs and Estimates			
Outcome	Mean	Standard Deviation	Adjusted R-Squared	85th Minus 50th Percentile of VAM Estimates (in z-score units)	Mean Standard Error (in z-score units)	Percentage of VAM Estimates that Are Statistically Distinguishable from the Average
Math PSSA, Grade 4	1,473	221	0.66	0.20	0.04	64.4
Math PSSA, Grade 5	1,477	223	0.76	0.19	0.04	66.4
Math PSSA, Grade 6	1,499	238	0.78	0.20	0.04	72.0
Math PSSA, Grade 7	1,500	240	0.80	0.18	0.03	72.6
Math PSSA, Grade 8	1,452	221	0.80	0.15	0.03	67.4
Math PSSA, Grade 11[a]	1,402	256	0.71	0.21	0.04	68.3
Reading PSSA, Grade 4	1,386	213	0.67	0.14	0.04	56.8
Reading PSSA, Grade 5	1,350	210	0.71	0.16	0.04	54.2
Reading PSSA, Grade 6	1,397	222	0.73	0.12	0.04	52.5
Reading PSSA, Grade 7	1,426	214	0.74	0.13	0.04	59.7
Reading PSSA, Grade 8	1,517	248	0.74	0.11	0.03	58.3
Reading PSSA, Grade 11[a]	1,399	260	0.67	0.17	0.05	58.1
Writing PSSA, Grade 5	1,339	256	0.50	0.31	0.05	75.8
Writing PSSA, Grade 8	1,410	262	0.54	0.29	0.05	75.3
Writing PSSA, Grade 11[a]	1,535	281	0.48	0.30	0.06	67.9
Science PSSA, Grade 4	1,461	182	0.64	0.24	0.05	69.0
Science PSSA, Grade 8	1,322	199	0.73	0.18	0.04	71.4
Science PSSA, Grade 11[a]	1,255	92	0.68	0.23	0.05	69.9
Holding Power, Grade 9[a]	91	25	0.40	0.58	0.05	92.2
Holding Power, Grade 10[a]	91	24	0.41	0.60	0.06	92.0
Holding Power, Grade 11[b]	88	33	0.73	0.93	0.05	90.0

Source: Mathematica calculations based on Pennsylvania student data.

Note: Unless otherwise noted, findings are based on three-cohort models with statewide samples of schools, principals, and students and a 95 percent confidence interval. The sample of principals consists of those who served as principals in every year from 2008-2009 to 2010-2011. One z-score unit is equal to one standard deviation of student outcomes.

[a] Findings are based on a two-cohort model.
[b] Findings are based on a one-cohort model.

Table C.6. Sample Characteristics of Outcome Measures and School VAMs Based on Phase 1 Samples

	Distribution of Student-Level Outcome Variable		Characteristics of School VAMs and Estimates			
Outcome	Mean	Standard Deviation	Adjusted R-Squared	85th Minus 50th Percentile of VAM Estimates (in z-score units)	Mean Standard Error (in z-score units)	Percentage of VAM Estimates that Are Statistically Distinguishable from the Average
DIBELS (NWF), Grade 1[a]	39	22	0.36	0.28	0.11	53.8
DIBELS (PSF), Grade 1[a]	40	15	0.59	0.38	0.09	61.5
DIBELS (ORF), Grade 2[a]	84	37	0.79	0.11	0.07	13.3
Math Progress Assessment, Grade 2[b]	73	16	0.64	0.23	0.08	66.7
Writing Progress Assessment, Grade 1[b]	76	14	0.47	0.40	0.10	53.8
Writing Progress Assessment, Grade 2[b]	76	15	0.53	0.26	0.09	53.8
Writing Progress Assessment, Grade 3[b]	75	19	0.50	0.16	0.10	21.4
Math PSSA, Grade 3[c]	1,298	188	0.64	0.23	0.07	37.5
Reading PSSA, Grade 3[c]	1,284	158	0.69	0.18	0.07	37.5
Attendance Rate, Grades 4-12[d]	93	6	0.48	0.08	0.10	13.8

Source: Mathematica calculations based on Pennsylvania student data and student data from pilot districts.

Note: Findings are based on a 95 percent confidence interval and one-cohort models with samples of schools, principals, and students from the pilot districts. The sample of principals consists of those who served as principals in 2010-2011. One z-score unit is equal to one standard deviation of student outcomes.

[a] Findings are based on students and teachers in Allentown and Cornell.
[b] Findings are based on students and teachers in Allentown.
[c] Findings are based on students and teachers in Allentown, Cornell, and Mohawk.
[d] Findings are based on students and teachers in Allentown, Mohawk, and Northwest.

DIBELS = Dynamic Indicators of Basic Early Literacy Skills; NWF = nonsense word frequency; ORF = oral reading fluency; PSF = phoneme segmentation fluency; PSSA = Pennsylvania System of School Assessment; VAM = value-added model.

MATHEMATICA
Policy Research

www.mathematica-mpr.com

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

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