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

Executive Summary

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

Executive Summary

April 5, 2012

Stephen Lipscomb
Hanley Chiang
Brian Gill

MATHEMATICA
Policy Research

All statistics are calculated by Mathematica unless stated otherwise

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.

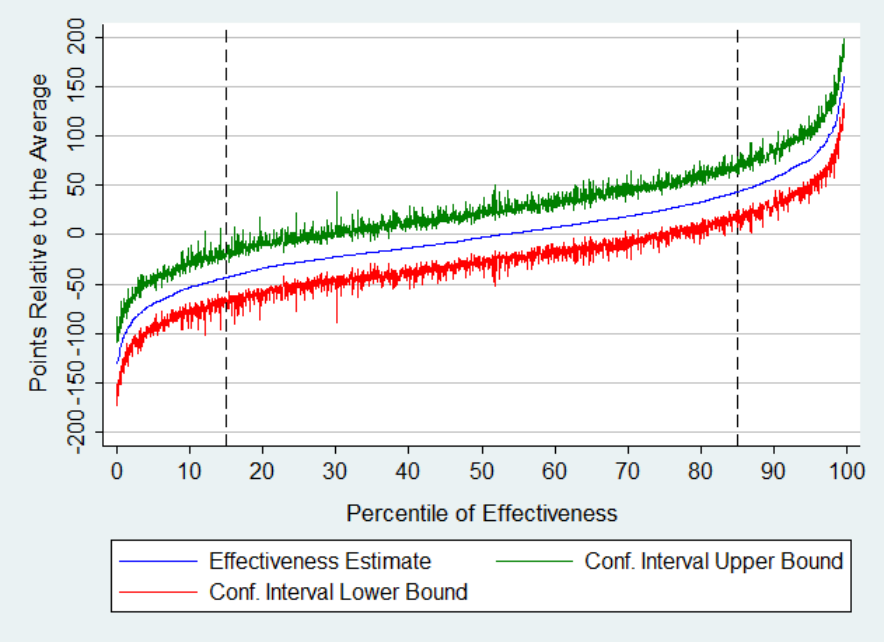
Using VAMs to Estimate Teacher Effectiveness

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.

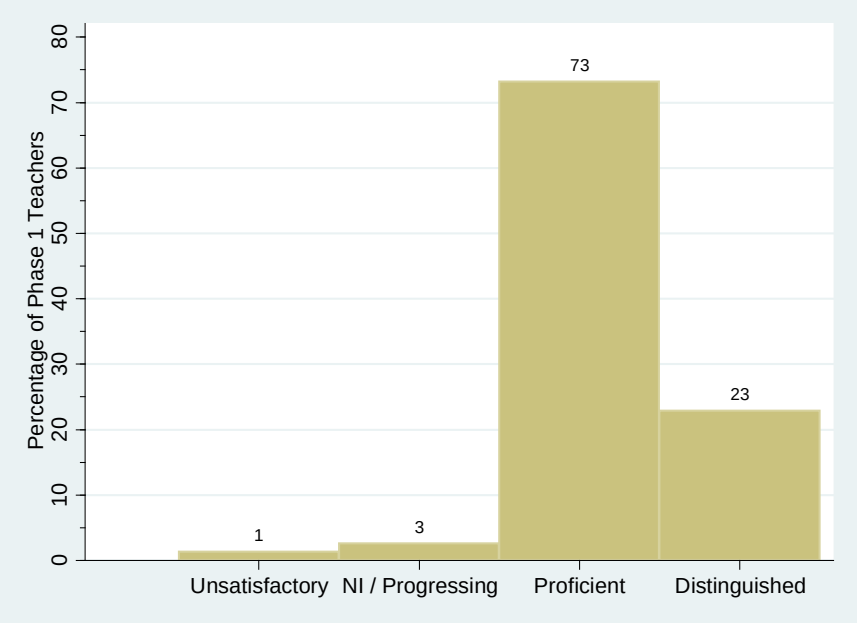
We estimated VAMs based on several non-PSSA outcomes and found notable differences in the ability of the VAMs to make reliable distinctions among teachers. For example, whereas 38 percent of teacher effectiveness estimates can be statistically distinguished from the average on the basis of a 1st-grade 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.

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