Teacher and Principal Value-Added: Research Findings and Implementation Practices

Final Report

September 14, 2010

Stephen Lipscomb
Bing-ru Teh
Brian Gill
Hanley Chiang
Antoniya Owens

MATHEMATICA
Policy Research, Inc.

THIS PAGE LEFT BLANK FOR DOUBLE-SIDED PRINTING

Mathematica Reference Number:
06815.100

Submitted to:
Dumaresq Associates
2090 Wexford Ct.
Harrisburg, PA 17112
Project Officer: Carolyn C. Dumaresq

Team Pennsylvania Foundation
100 Pine Street, 9th Floor
Harrisburg, PA 17101
Contract Officer: Matthew A. Zieger

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

Teacher and Principal Value-Added: Research Findings and Implementation Practices

Final Report

September 14, 2010

Stephen Lipscomb
Bing-ru Teh
Brian Gill
Hanley Chiang
Antoniya Owens

MATHEMATICA
Policy Research, Inc.

THIS PAGE LEFT BLANK FOR DOUBLE-SIDED PRINTING

CONTENTS

THIS PAGE LEFT BLANK FOR DOUBLE-SIDED PRINTING

TABLES

EXECUTIVE SUMMARY

Background on Value-Added Models

In recent years, policymakers, educators, and researchers have worked to better understand and measure the effects of teachers and principals on student achievement. Previous performance measures based on student achievement data relied largely on average *levels* of achievement. Now, however, there is a shift towards using grade-to-grade *growth* trajectories of individual students to produce better estimates of what educators contribute to achievement, regardless of where their students start. These models, known as value-added models (VAMs), are based on predictions about each student's expected achievement in a subject and grade were s/he to have the "average" teacher.[1] The predictions are a function of previous achievement and often other student characteristics. VAMs then measure the extent to which actual achievement in a classroom or school differs above or below the predicted level of achievement. A central principle of value-added modeling is that teachers are not held responsible for their students' incoming achievement, but rather are evaluated by how much they contribute to their students' learning.

Mathematica's Review of Research Findings and Implementation Practices

This report summarizes research findings and implementation practices for teacher and principal VAMs, as a first step in the Team Pennsylvania Foundation's (Team PA) pilot project to inform the development of a full, statewide model evaluation system. We have selected 21 studies that represent key issues and findings in the literature and examined varying degrees of value-added implementation in seven school districts or states. We present information aimed at VAM development: typical data elements, important modeling considerations, features specific to teachers or principals, and broad implementation features.

Though analysts' models and methods differ, their findings consistently indicate that students' prior academic histories have by far the most explanatory power among factors that predict student performance in a given year. Nonetheless, findings also consistently confirm that highly effective teachers have a meaningful impact on student achievement growth. In contrast, far less evidence is available on the effects of principals on student achievement: Nearly the entire value-added research base examines impacts for teachers. There is very little research that attempts to separate the principal effects from those of teachers or schools. The only study we found that attempts to identify and study individual principal effects separately from teacher or school effects suggests that principals have a meaningful effect on student achievement growth as well.

Applicability of VAMs to Teachers across Grades and Subjects

Scholarly articles have examined a narrower range of grades and subjects than stakeholders would prefer—commonly math and reading in grades 4 to 8. VAMs used by policymakers in areas such as Tennessee and Dallas include more grades and subjects than in the literature, and they also apply value added to principals. They take advantage of multiple available assessment measures, including not only the state standardized assessment used for federal accountability purposes but also other assessments, such as local end-of-course assessments. But all the VAMs we examine—in research literature and in practice—rely on student test scores. Test scores are available and

[1] Our definition of VAMs includes all statistical models that produce estimates of the effects of individual educators (teachers, principals and/or schools) on any outcome that they might plausibly affect—not just test scores.

quantifiable for students in multiple grades, and most possess some good measurement properties. We have not encountered any examples of VAMs that use non-assessment outcomes, such as student attendance. As a result, value-added is not estimated for educators who teach subjects or grades without some kind of standardized test outcome. Policymakers sometimes re-weight other evaluation components or substitute a school-wide VAM estimate for educators for whom individual value-added cannot be calculated. Applying VAMs broadly across grades and subjects that are untested remains an important challenge for both researchers and practitioners.

Producing Valid and Reliable VAM Estimates

Much of the research literature on VAMs focuses on ways to produce estimates that are unbiased and estimates that have sufficient reliability. VAMs vary somewhat in the methods they use in aiming to produce unbiased estimates (that is, valid estimates of educators' contributions to student learning). For example, some VAMs control for external effects on student learning by relying entirely on prior information about each student's achievement; others incorporate statistical controls for student demographic characteristics alongside controls for prior achievement. In addition, VAMs for teachers differ in whether they subtract estimated effects of school-wide performance to provide fuller statistical controls for unobserved school-level student, neighborhood, and school factors that are external to teacher contributions in the classroom.

All estimates, even unbiased ones, have some degree of random variation or statistical "noise." Noise originates from the chance that individual students in a teacher's classroom or in a principal's school perform unexpectedly well or poorly. This noise tends to average out over a large number of students but makes inferences less reliable when based on only a small number of students. Analysts have considered and incorporated several modeling features aimed at reducing random variation—that is, increasing the precision of the estimates. Some of these adjustments also help to ensure that teachers with fewer students are not overrepresented among those at the extremes of the value-added distribution merely because of unexpectedly high or low performance by a small number of students. Analysts have examined the reliability of estimates by exploring whether high value-added teachers tend to be high performing in successive years. Findings typically show a statistically significant, albeit moderate, amount of year-to-year correlation. In addition, researchers have pointed out that including multiple successive cohorts of a teacher's students (that is, averaging estimates over two or more years) can reduce noise, improving precision as a result of having more data.

Both the number of years of teaching included in VAM estimates and the inclusion or exclusion of school-wide effects in teacher estimates are examples of how VAM development involves policy decisions for stakeholders that must be informed by their goals and by research. Including multiple years of teaching data improves the reliability of estimates, but it also makes the estimates less timely and makes it difficult to assess recent changes in performance. Likewise, adding school-wide effects to teacher VAMs arguably provides better statistical controls but it implicitly creates a comparison of teachers within each school, thereby potentially undermining incentives for cooperation within schools (depending on how the VAM estimates are used and what stakes are attached).

Using Value-Added and Measures of Professional Practice Together

Recent findings indicate that teachers' value-added estimates correlate with measures of professional practice, such as observed classroom behavior and principals' ratings of performance. These findings suggest that, by using value-added estimates in conjunction with observational

measures, we can learn not only which educators are highly effective but also how their teaching practices enable their success.

The validity of a VAM as a broad measure of teacher or principal performance is limited by the extent to which the assessment measures everything that students should learn and teachers should teach. Policymakers typically include value-added estimates as one component in teacher or principal evaluation models (when value-added is used at all), along with measures of professional practice. Component weights vary across systems, with value-added being a relatively small factor in some systems to half of the overall evaluation in others. The push to incorporate information from VAMs has grown in recent years, particularly as a result of the U.S. Department of Education's Race to the Top competition, but at the moment there is little consensus among scholars about what the optimal weights for evaluation components should be. Weighting is partly, and perhaps primarily, a policy question, but it can be informed by empirical analysis.

Principal Value-Added Models

The principal VAMs that we have seen in the research literature and in practice (for example, in Tennessee and Dallas) generally assume that the principal's value-added is the same as the school's value-added. The school VAM examines the extent to which actual student achievement levels at the end of the year in a school differ above or below predicted achievement, which essentially averages teacher value-added scores across teachers. The school model is applied to principals explicitly in Tennessee and Dallas, and implicitly in research models that do not separate the contributions of principals from the contributions of schools. This type of VAM has the advantage of being estimable for all principals. However, it may not be the most accurate way to represent principal contributions because principals affect student achievement in different and less direct ways than do teachers at the school.

Our concern is that, just as a teacher should not be penalized for serving disadvantaged students with low achievement levels when he or she enters the classroom, a principal should not be penalized for taking on the leadership of a school that has chronically underperformed in the past. In other words, the *improvement* in school value-added scores from one year to the next may be a better way to estimate principal value-added. Analysts are also considering models that do separate the contributions of schools from the contributions of principals. While these models arguably provide estimates that are more valid, they rely on principal turnover and mobility to separate principal and school effects and therefore cannot provide timely estimates for all principals. The best way to estimate principal value-added comprehensively and rigorously is not yet clear.

Discussion

Overall, the findings from our review suggest that VAMs can provide meaningful, if noisy, information about teacher effectiveness. Much less is known about principal effectiveness. We caution against assuming that models developed for other subjects, grades, and personnel can necessarily be extended to other applications. For example, our exploratory analyses of Pittsburgh high schools suggest that student attrition can lead to substantial bias; that is, schools in which large numbers of students with low growth trajectories drop out appear highly effective in the data because the models include only the remaining students. The bias is particularly severe when the nearest available baseline score is several grades earlier, as it is in Pennsylvania when using statewide assessment scores from grades 8 and 11. This problem illustrates how clear technical and practical challenges remain, and how VAM development requires care to produce results with the best statistical properties. But the methods and applications are also improving at a fast pace, and

Mathematica teams are involved in several of these efforts. We look forward to working with Team PA and the stakeholder steering committee throughout this pilot study to help develop state-of-the-art VAMs that serve their intended purposes and are valid and reliable.

INTRODUCTION

The empirical evidence on the contributions of educators to student achievement growth has expanded considerably in recent years. The combined interests of stakeholder groups that include policymakers, educators, and researchers have fueled these advances. A particular focus has been to use the year-to-year growth trajectories of individual students to produce valid estimates of what educators are contributing to achievement, regardless of where their students start. By their construction, these estimates can overcome a main deficiency of many other measures in regular use today—such as average score performance or the rate of student proficiency—that can penalize teachers and principals for serving disadvantaged student populations. The statistical methods used to estimate educators' contributions to student achievement, known as value-added models (VAMs), vary in their details but have in common a general approach that relies on information about individual students' achievement in other years and/or academic subjects.

VAMs are based on predictions about each student's expected achievement in a subject and grade with the "average" teacher.[2] These predictions use previous achievement and often other student characteristics that are related to student achievement growth but are outside the control of teachers and schools. A teacher VAM, for example, examines the extent to which the actual achievement of a teacher's students at the end of the year is above or below their predicted achievement. A principal VAM analogously would measure improvement in the performance of the school he or she serves compared with the prediction for the "average" principal. The capacity to conduct value-added analyses relies foremost on district or state data systems that link students to their teachers, schools, and prior achievement histories. Although the models are still being refined, value-added is already recognized as having the potential to provide better information than ever before on the effectiveness of individual educators in raising student achievement.

This report summarizes current research findings and implementation practices on value-added modeling for teachers and principals, as a first step in a pilot project to inform the development of a full, statewide model evaluation system in Pennsylvania. We examine the features of several VAMs in use today and synthesize results from more than 20 recent studies that span important strands in the research literature to aid discussions among stakeholders involved with VAM development.[3] These studies do not comprise the entire VAM literature, but they are representative of key issues and findings. Our reviews highlight six general points about the literature and application of value-added measures.

1. *Highly effective educators have a meaningful impact on short-term student achievement growth.* Data from several cities and states suggest that the top 15 percent of math teachers are capable of raising the achievement of the median-performing student by about 5 to 8 percentile points with one year of teaching (adjusting for student characteristics).[4] By the

[2] VAMs are inherently norm-based. Teacher or principal effects are measured relative to the effectiveness of the average teacher or principal, respectively.

[3] We limit our scope to studies published or developed since 2005. McCaffrey et al. (2004) provide an excellent review of earlier value-added studies.

[4] This comes from the standard deviation of teacher value-added estimates, which is a measure of their variability. The amount of variability in value-added estimates across teachers relates to the size of the relative benefit of a high value-added teacher opposed to a low value-added teacher. Were the distribution narrower than it is measured to be, the effect of a high value-added teacher would be less because teachers would be grouped close together around the average. See Appendix A for details on each study.

same token, the bottom 15 percent of math teachers appear to lower the median student's performance by the same amount. Variability in value added for reading teachers and principals also appears to be large (though the data on principals comes from one study).

2. *Nearly all value-added research studies examine applications to teachers and schools, but not principals.* Only now are analysts starting to consider principal VAMs in earnest. Although many of the methodological considerations that apply to teacher and school VAMs extend to principal VAMs as well, there are important distinctions too. A principal's value added should not be presumed to be synonymous with a school's value added (or the average value added across teachers) because principals affect student achievement growth in different and less direct ways than teachers do.

3. *Current VAMs measure student growth exclusively through assessment scores.* Assessment data are available in multiple grades (to account for student growth trajectories) and typically have some good measurement properties (to help alleviate concerns that the measurement of the outcome will lead to biased or unreliable inferences about effectiveness). Analysts and practitioners have thus far limited VAMs to core subjects and to grades for which current and prior assessment data are available. Identifying accurate and reliable ways to expand VAMs to nontested subjects and to early elementary and all high school grades remains a challenge. Moreover, even in tested grades and subjects, the validity of a VAM as an estimate of an educator's performance depends on the extent to which the student assessment captures the range of skills and knowledge that students are expected to learn. For example, if the student assessments do not capture students' higher-order thinking skills, or their ability to write a coherent essay, then the VAM estimates will not capture teachers' contributions to these aspects of learning.

4. *VAMs vary in how they address potential bias in estimates.* Many scholarly articles on VAMs focus on ways to produce unbiased estimates (that is, estimates that are not systematically overrepresented or underrepresented). A common strategy is to include controls for any available student or classroom characteristic to better isolate the contributions of educators themselves. Another strategy is to include more years of student assessment results as control variables. Some studies also subtract out the average effect of each school. This approach can reduce bias due to omitted school variables and address the concern that teachers and students are sorting into schools nonrandomly. However, it does not allow for cross-school comparisons because it essentially compares teachers to the "average" teacher within the same school only.

5. *Like all estimates, value-added estimates have some degree of random variation or statistical "noise."* Ensuring sufficient precision is another focus in the research literature on VAMs. For example, analysts have shown that averaging teacher performance across two or more years can reduce year-to-year errors. But even well-specified models that have substantial predictive ability will not be totally free of noise. When producing VAM estimates, many researchers include explicit estimates of noise by indicating a confidence interval, or range of likely true value added, around each estimate. VAMs also now regularly employ a method that helps to reduce the chance that teachers in small classes are overrepresented at the tails of the value-added distribution because of the performance of one or two students.

6. *Value added can be used with professional practice measures to help educators improve performance.* Value-added estimates are inherently "black box" descriptions of effectiveness: They will, ideally, help to identify teachers and principals producing especially high (and low)

levels of achievement growth in their students, but VAM estimates cannot by themselves determine how or why some teachers and principals are producing more achievement growth than others. Analysts are finding that VAM estimates correlate with measures of professional practice, suggesting that the estimates might be measuring similar teacher attributes.

The following sections of this report summarize findings from our literature review. The body of the report considers, in turn, main findings from the research literature that help inform stakeholders about best practices in VAM methods (such as typical data elements, important modeling considerations, and the specific application of value added to principals) and broad implementation features of several existing VAMs. Appendix A provides details on each study we examined. Appendix B provides details on implementation features in several school districts and states. Finally, Appendix C provides a brief technical description of value-added models.

RESEARCH FINDINGS FOR INFORMING VAM DEVELOPMENT

A. Components of a VAM

A typical VAM includes several components: an outcome measure, a baseline (or prior year) measurement of the outcome, control variables, and teacher or principal variables. Depending on the context for which the VAM is used and the availability of data, the researcher might elect to use different sets of outcome and control variables. Below, we discuss some of the choices that researchers have made in the literature.

1. Outcome Measures

Each of the VAMs we reviewed used at least one standardized test score as the outcome measure of interest. For example, the Tennessee Value Added Assessment System (TVAAS) includes student scores in grades 3 to 8 on the Tennessee Comprehensive Assessment Program, a series of assessments in reading, language arts, math, science, and social studies. At the high school level, it also includes statewide end-of-course tests in Algebra I, Biology, and English II. Dallas's Classroom Effectiveness Index (CEI) includes scores in a similar set of subjects as measured by both the Texas Assessment of Knowledge and Skills and by lower-stakes assessments (that is, assessments that are not used for state accountability purposes).

There appears to be no consensus among researchers as to whether high- or low-stakes assessments are preferable. The choice of assessments might be driven primarily by the availability of data, rather than by concerns over issues such as whether a given assessment captures effects related to teachers "teaching to the test" or the extent to which students have incentives to perform well. For example, analysts in 8 studies used scores from a low-stakes test; 11 used scores from a high-stakes test;[5] and 3 others used scores from unidentified tests. McCaffrey et al. (2009) observe that the year-to-year correlations of value-added estimates differ when using scores from low- versus high-stakes tests, although no clear pattern emerged across the five large Florida districts they examined. Part of the reason why the effect of existing stakes might not be resolved is that estimating and reporting value added for individual educators implicitly raises the stakes for any test that is used as the outcome—and this raising of stakes becomes explicit if the VAM estimate is used for teacher evaluation or pay.

[5] Two studies used scores from both a low- and high-stakes test.

Analysts have also examined whether the construction and scaling of test scores affects VAM scores. For example, score ceiling effects (that is, when students can earn the maximum score and therefore appear to have no room for academic improvement) have the potential to introduce bias in VAM estimates. Findings suggest that most ceiling effects that are typically observed have a negligible influence on teachers' VAM scores. However, when analysts imposed an artificial ceiling on scores at a very low level, such as would be the case for minimum competency measures, they found that relative rankings of teacher VAM estimates shifted considerably (Koedel and Betts 2010).

None of the studies we examined used outcome measures, such as student attendance or credit completion, other than test scores in VAMs to estimate teacher effectiveness.

2. Control Variables

Baseline test scores are the most important control variables to include in a VAM because they are the strongest predictors of current and future test achievement. Twenty studies used at least one year of baseline test scores as controls. Although studies indicate that controlling for more years of test score history helps to reduce bias in VAM scores due to nonrandom sorting of students into classrooms (Rothstein 2009;[6] Koedel and Betts 2009), only six studies use two or more years of baseline test scores as controls. Seven[7] studies use only the baseline test score of the same subject as the outcome measure as a control; 11 studies also use baseline test scores from other available subjects as controls.

Dallas's CEI system and all but two of the VAMs we reviewed from the research literature also include controls for student characteristics. There is a compelling argument that one should control for everything that the teacher cannot affect, so recorded student characteristics are included in the models especially when it is virtually costless to do so. The most common student controls are for gender, race/ethnicity, disability/special education status, free or reduced-price lunch status, and English language proficiency level. Some studies also controlled for parental education, number of hours of television watched during the week, family income, and so on. It appears that the availability of data largely determines the set of student characteristics included in the VAM. Some models, however, assume that all this information is subsumed in the baseline scores already. VAMs based on the model by William Sanders (that is, TVAAS and similar models in Pennsylvania and North Carolina) do not include controls for student characteristics. Ballou (2005) finds that omitting student characteristics in the TVAAS model that accounts for up to five years of test score growth trajectory does not appear to suffer from substantial bias.

Another commonly included group of controls are peer/classroom-level characteristics, such as class size and classroom averages of the student characteristics described in the previous paragraph. These controls help to separate peer effects in the classroom from the contributions of individual educators. A smaller number of studies (four) also included time-varying school-level characteristics such as school size, average class size, and percentage of teachers with 10 years or more experience.

[6] Rothstein (2009) analyzes bias by assuming that the "true" teacher model includes three years of prior scores in math and reading and student characteristics. He then estimates the amount of bias relative to the true model for models that include (1) one year of prior scores in both subjects and no student characteristics and (2) three years of prior scores in both subjects and no student characteristics. The first comparison shows much less bias than even less sophisticated models (16 percent of the total variance is bias). The second comparison shows almost no bias (4 percent of the total variance). All models in this study include school effects as well.

[7] One of the studies incorporated the baseline test score into the outcome variable by using the gain between the current and previous years' scores as the outcome variable.

Approximately half the studies subtract the average estimated effect of the school (that is, they incorporate school fixed effects) in estimating teacher effects, either in the main VAM or in secondary VAMs. This method can help reduce omitted variable bias and address concerns related to nonrandom sorting of teachers and students into schools. An important consequence of including these school-level controls is that it implicitly assumes that the average teacher in one school has the same value added as the average teacher in every other school. Depending on the application of the VAM estimates, it might be less desirable to include school effects. For example, when trying to identify the lowest- and highest-performing teachers in a district or state, school effects will have to be left out of the model. This is because VAMs measure effectiveness relative to the average among the teachers in the sample and including school effects will restrict the comparison only to teachers in the same school.

A couple of practical considerations could complicate the effort to attribute achievement growth to the correct teacher. The first of these is team teaching, whereby each member of the team is responsible for teaching a certain set of topics to all students within that particular grade or subject. Also, students who change schools within or across districts during the school year make it difficult to attribute achievement growth among the different teachers who were responsible for the student at different times during the year. A possible solution is to use a dosage approach to account for the amount of time each student spends with a teacher during the school year (Ballou 2005; Lipscomb et al. 2010). The rest of the studies we examined did not use dosage measures and instead included only a binary variable for whether a student was taught at all by a teacher each year.

B. Modeling Considerations

Given the outcome variables, covariates, and teacher dosage variables that are selected for inclusion in a VAM, the relationships among these variables must be rigorously modeled. In prior studies and applications of VAMs, analysts have considered and incorporated several modeling features aimed at producing VAM estimates with desirable statistical properties. The most important features observed in previous VAM designs have been motivated by two central objectives. First, analysts have sought to minimize *estimation bias*, or the extent to which teachers' effects are systematically overestimated or underestimated. Second, analysts have sought to ensure sufficient *precision*—that is, a sufficiently small margin of error around the teacher effect estimates.

Next we discuss four key modeling features that have been prevalent in prior value-added analyses: (1) selecting the number of years of teaching data on which VAM estimates are based, (2) adjusting estimates on the basis of their precision, (3) accounting for test measurement error, and (4) handling missing data. All four features have been motivated by considerations of bias and precision. We then describe ways in which analysts have assessed the precision of VAMs by examining the consistency of VAM estimates for a given teacher.

1. Selecting the Number of Years of Data Used in Estimation

An important modeling choice is to select the number of years of student growth data on which each teacher's VAM estimate is based—that is, the number of current and prior student cohorts who contribute to a teacher's current VAM estimate. This choice entails a balance between allowing VAM estimates to be reflective of teachers' most recent performance and enhancing the precision of the estimates by employing multiple years of data (Goldhaber and Hansen 2008, 2010; Lipscomb et al. 2010). Conceptually, these precision gains stem from the fact that with more years of data larger sample sizes of students are used to produce a teacher's VAM estimate; this dampens the random fluctuations in the estimates that stem from being assigned, by chance, a few students with unusually

high or low learning growth (McCaffrey et al. 2009). Because multiple-year VAM estimates are less prone to these random errors, there is less statistical uncertainty in identifying performance differences among teachers on the basis of these estimates.

Much of the existing literature on choosing the sample duration has focused on quantifying precision differences between multiple- and single-year VAM estimates. One common approach is to compare the percentage of teachers whose performance estimates are statistically distinguishable from average performance in the sample under a one-year VAM and a three-year VAM. When switching from a one-year to a three-year model, the reported increase in the percentage of teachers with statistically distinguishable estimates varies widely across studies, from a small increase of 3 percentage points to a substantial gain of 28 percentage points (Ballou 2005; Goldhaber and Hansen 2008; Lipscomb et al. 2010). Nevertheless, the consistent finding is that multiple years of data yield a greater ability to detect performance differences.

There is also some evidence that averaging VAM estimates across multiple years can reduce estimation bias stemming from *systematic*—and not just chance—assignment of unusually high- or low-growth students to particular teachers. Using data from a single statewide cohort of students, Rothstein (2010) finds that specific fifth-grade teachers are more likely than their colleagues—to a degree not consistent with pure chance—to be assigned students who demonstrated particularly high or low gains in fourth grade. Although Koedel and Betts (2009) find similar results using data from a different school system, they document that the sorting of students into particular teachers' classrooms on the basis of learning growth trajectories is much less pronounced when three to four cohorts of students are combined in the VAM analyses. Thus, even systematic sources of bias can offset each other when VAM estimates are averaged across years.

2. Adjusting VAM Estimates on the Basis of Their Precision

In addition to reporting measures of the precision of VAM estimates, it has been common for value-added analyses to incorporate an adjustment to the estimates that directly reflects their level of precision. This adjustment, known as empirical Bayes estimation or shrinkage, is motivated by a simple fact: 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 (such as an illness on the test day or an unusual familiarity with the topic of a reading passage). In the absence of further adjustment, teachers with fewer students—that is, those with less precise estimates—will be overrepresented in the extreme portions of the estimated performance distribution (at both the high and the low end) due purely to larger fluctuations from these chance factors (Lipscomb et al. 2010).

Shrinkage adjustments account for the fact that estimates with greater precision carry greater strength of information about teachers' true performance levels. A teacher's adjusted estimate is a weighted average of his or her own initial estimate and the mean estimate of all teachers in the sample, with more precise initial estimates receiving greater weight (Jacob and Lefgren 2008; Kane and Staiger 2008). In essence, teachers are assumed to be average in performance until evidence accumulates to justify a different conclusion. The conceptual appeal of this adjustment is reflected in its prevalent application: among the studies reviewed that estimate teacher VAMs, we are able to confirm that at least 10 apply the shrinkage approach.

In addition, to further minimize the risk of making erroneous conclusions on the basis of imprecise estimates, half of the reviewed studies also limit analyses to teachers who have taught at least a specified minimum number of students. The specified minimum, typically chosen in an ad

hoc manner, has varied from 5 students (Harris and Sass 2009) to 20 students (Koedel and Betts 2009, 2010).

3. Accounting for Test Measurement Error

A central principle of value-added modeling is that teachers are not held responsible for their students' incoming achievement, but rather are evaluated by how much they contribute to their students' learning after they enter the particular teacher's class. It is thus important for VAMs to control accurately for differences across classrooms in students' prior achievement. However, measurement error in pretest scores—due, for instance, to random variation in how students feel on testing day—obscures some of the true differences in students' prior achievement that ought to be controlled.

Previous work has employed various methods to control more accurately for students' prior achievement in the presence of pretest measurement errors. One straightforward approach is to include multiple measures of students' prior achievement in the set of control variables. Nine of the VAM studies in Appendix A use information from multiple pretest scores in the analysis. In most of these cases, the VAMs control for scores from multiple subjects in the previous year only; however, some analysts have considered specifications that use information from multiple years of prior scores (Ballou 2005; McCaffrey and Hamilton 2007; Rothstein 2009, 2010; Lipscomb et al. 2010). Alternative approaches to addressing pretest measurement error include using instrumental variables methods in which only the pretest score variation unrelated to measurement errors is used in VAM estimation (Potamites et al. 2009) and using data on the precision of the test to adjust the results (Rothstein 2009).

4. Handling Missing Data

It is common for at least some student records in a data system to be incomplete—that is, to have missing values for one or more outcome or control variables used in the VAM or to be missing a link to the appropriate teacher whose VAM estimate should reflect the student's growth. Prior VAM analyses have handled these missing records in a number of ways. The vast majority of VAM analyses exclude student records with missing values for outcome or baseline score variables, although models such as TVAAS include records with missing baseline scores. Only a handful of studies have sought to retain incomplete records in the analysis through statistical methods that impute missing data. Imputation methods used by these studies have included maximum likelihood methods (McCaffrey and Hamilton 2007) and Bayesian methods (Mariano et al. 2010). Occasionally, analysts have employed different methods for different types of variables. Although Potamites et al. (2009) exclude records with missing posttest or pretest scores, they replace missing values of demographic control variables with regression-predicted values. In general, for the most important variables in value-added analysis, the posttest and pretest scores, the value-added literature has not converged on a single, widely accepted method for incorporating records with missing values.

5. Evaluating the Year-to-Year Consistency of VAM Estimates

In addition to incorporating technical features into the model design aimed at minimizing bias and improving precision, analysts have also assessed the statistical properties of the actual estimates generated by the VAMs. Among the most common types of analyses are those that evaluate the consistency of VAM estimates for the same teachers over time. Because true teacher performance can change over time, VAM estimates for a given teacher are not expected to be perfectly consistent across time periods. Nevertheless, there should be some degree of consistency if the VAM estimates

are accurately capturing at least some permanent component of a teacher's effectiveness. In contrast, if VAM estimates primarily reflect transitory estimation error, then a teacher's estimate in one period will have little association with that in another. Regardless, the consistency of VAM estimates depend in part on the reliability of these estimates—that is, the extent to which measured performance differences reflect persistent differences in effectiveness rather than error-induced imprecision or time-varying performance.

Prior work has gauged the consistency of VAM estimates in various ways. Several studies have calculated the correlation between teachers' single-year VAM estimates from consecutive years; these correlations have generally been moderate in magnitude, ranging from 0.2 to 0.7 (Goldhaber and Hansen 2008, 2010; Hanushek and Rivkin 2008; Lipscomb et al. 2010; McCaffrey et al. 2009). A related measure is the extent to which teachers' performance rankings change from year to year; a number of studies have found that about one-third to two-fifths of teachers in the top or bottom grouping of performance ranks (typically, a quartile or quintile) stay in the same grouping in the subsequent year (Aaronson et al. 2007; Ballou 2005; Lipscomb et al. 2010; McCaffrey et al. 2009). A third approach has been to estimate regressions to determine whether teachers' VAM estimates in an earlier period can significantly predict the achievement gains of (different) students taught by these teachers in later periods; teachers' earlier VAM estimates have been consistently found to be substantively and statistically significant predictors of their impacts on later student cohorts (Goldhaber and Hansen 2010; Harris and Sass 2009; Kane and Staiger 2008; Rockoff and Speroni 2010). In fact, Kane and Staiger (2008) find that differences in prior VAM estimates within pairs of teachers closely replicate differences in experimentally estimated—and thus unbiased—teacher impacts within these same pairs in subsequent years. In sum, the literature has uniformly found that teacher VAM estimates should exhibit at least a moderate degree of consistency over time.

C. VAMs for Principals

Nearly all research conducted on VAMs has focused on applications to teachers and schools; few scholarly articles have attempted to estimate the performance of individual principals. Principals affect student achievement differently and less directly than teachers do, such as through effective organizational management, by recruiting and retaining effective teachers, and by ensuring a working environment in which teachers can be effective. Although many of the same practical and theoretical issues apply, we caution against assuming that a principal's value added is necessarily the same as a school's value added (or necessarily the same as an average value added across teachers).

The only paper we found that applies a VAM specifically for estimating individual principal effectiveness is a working paper by Branch et al. (2009)[8] from the 2009 Center for Analysis of Longitudinal Data in Education Research (CALDER) conference. The authors study principals in Texas with two to three years of principal experience at their respective schools. As in teacher VAMs, Branch et al. use shrinkage estimates and are concerned about the possibility of bias due to nonrandom student sorting into schools. They describe a reweighting procedure that helps the VAM address this type of bias by treating principal contributions equally whether they have mostly high- or low-achieving students. If student observations are not reweighted, principal efforts at schools with primarily high-achieving students might appear less meaningful because students start out at a high level already.

[8] As such, please note that this paper is preliminary and findings are subject to change.

The models they consider measure effectiveness using reading and math scores from the Texas Assessment of Academic Skills (TAAS) and control for a single year of baseline test scores, student characteristics, time-varying peer and school characteristics, year-by-grade indicators. Using this framework, Branch et al. then estimate principal effectiveness in two ways. The first model includes principal-by-school effects (rather than teacher effects) along with the variables listed above. The second model includes school effects along with the principal-by-school effects.

The findings in both cases are norm-based—as they are for teachers. The first model compares principals across schools in Texas and has the advantage of being estimable for all principals in whatever sample is used. The disadvantage is that value-added measures based on this model are likely to combine the contributions of principals with the contributions of other staff at the school. The formulation of the model closely resembles a model that analysts have used for estimating school value added (Lipscomb et al 2010; Potamites et al 2009). It is also consistent with approaches that Dallas and Tennessee are using for measuring principal effectiveness, which assign school value added to principals directly (see following section). We plan to consider whether assigning school value added to principals is the most accurate way to represent principal contributions. Our concern is that, just as a teacher should not be penalized for serving disadvantaged students with low achievement levels when he or she enters the classroom, a principal should not be penalized for taking on the leadership of a school that has been chronically low performing, based on school value added, in the past. In other words, we will study whether *improvement* in school value added over time is a better way to measure principal effectiveness.

The second model that Branch et al. consider explicitly separates principal contributions from contributions of the school. The advantage of this approach is that it plausibly produces estimates that are more valid but the disadvantage is that it is considerably limited in its application across principals. The model compares principals over time at the same school, that is—it relies on principal mobility and turnover to separate the effects of principals from the effects of schools. As a result, principals who are the only principal at a school during the sample period cannot be included. The model also cannot isolate a principal's value added in any particular year.

As this tradeoff between internal and external validity illustrates, the best way to estimate principal value added comprehensively and rigorously is not yet clear. Interestingly, however, the authors' estimates of the variation in principal VAM estimates compares to the amount of variation that many studies observe for teacher effectiveness. The findings suggest that the benefit of having an effective principal might be large for students, particularly those in high-poverty schools.[9]

THE IMPLEMENTATION OF VAMS IN SCHOOL DISTRICTS AND STATES

Tennessee and Dallas were the first to use value added for measuring teacher and school effectiveness, with models that are more than a decade old. The push to consider such systems has expanded in recent years, particularly as a result of the U.S. Department of Education's Race to the Top competition. VAMs in use today vary in both their designs and their stakes for individual educators, ranging from providing administers with high-level data for professional development purposes to being important factors in yearly evaluations and recertification. Although momentum appears to be shifting toward considering the potential for VAMs in measuring teacher and principal

[9] The study also finds (1) that principals follow a similar pattern as teachers in terms of "preferring schools with less demands as indicated by higher income students, higher achieving students, and fewer minority students," and (2) that principal effectiveness improves slightly with tenure at a school.

effectiveness, proponents of implementing such systems have met with varied levels of success. In May 2010, Governor Bobby Jindal of Louisiana signed into law a bill that requires school districts to give a 50 percent weight to value-added estimates in yearly teacher evaluations.[10] One month earlier, Governor Charlie Crist of Florida vetoed an even stronger measure that would have given the same 50 percent weight to value-added estimates and eliminated tenure for all new hires.[11] In July 2010, D.C. Schools Chancellor Michelle Rhee dismissed 241 teachers, including 165 teachers who received poor appraisals under a new evaluation system that includes a component measuring teachers' effectiveness based on their students' performances in standardized tests.

We discuss implementation features of several existing or developing VAMs, focusing on the following: (1) how districts and states are using value-added information; (2) value added as a component in composite evaluation measures; and (3) the extent to which staff across grades and subjects are included. For this review, we examined systems in Dallas, Florida, Louisiana, Memphis, North Carolina, Pennsylvania, and Tennessee.[12] These school districts and states all have the necessary data systems to use value added in teacher and principal evaluations but do so to varying degrees. Appendix B provides additional information on each system.

1. How School Districts and States Are Using Value-Added Information

Four of the locations we examined—Dallas, Tennessee, Louisiana, and Memphis—use value added directly for helping to identify highly effective teachers and principals. The current Dallas model is an opt-in pay-for-performance system that provides awards of up to $3,200. Neither the decision to opt in nor the size of the performance award directly affects yearly evaluations for teachers and principals. Dallas' pay-for-performance system augments the district's longstanding use of value-added data for recognizing schools with high student growth rates. Tennessee now considers value-added data to be a standard component in teacher and principal evaluations. Though the state has provided value-added reports to schools and teachers for years, the state's successful bid for Race to the Top funds has ushered in an expanded use of value added for high-stakes purposes, such as determining compensation, promotion, retention, and tenure.[13] Similarly, Louisiana's newly enacted system will eventually require that school districts use value added directly in evaluating teachers and administrators. The state also plans to use value-added scores for recertification purposes: Teachers who are rated as ineffective at least three times in their certification cycle (based on an overall evaluation measure) will not be recertified unless school boards make a successful appeal.[14]

Florida, North Carolina, and Pennsylvania are examples of states that collect rich information on students and teachers but do not use VAMs for evaluating teacher or principal effectiveness. With the failure of the teacher evaluation bill in Florida, the state continues an evaluation system that limits its use of student achievement measures to those based on levels (that is, not relative to a baseline), such as performance on state and local assessments. Both North Carolina and

[10] http://www.nola.com/politics/index.ssf/2010/05/new_teacher_evaluation_system.html

[11] http://www.miamiherald.com/2010/04/16/1582150/why-charlie-crist-vetoed-the-teacher.html

[12] We consider Memphis separately from Tennessee, although Memphis City Schools are bound by the statewide system as well.

[13] See Tennessee's guide to its new teacher and principal evaluation system, available at http://www.tn.gov/firsttothetop/resources.html.

[14] The Louisiana teacher and principal evaluation system will be piloted in 2010–2011 and 2011–2012, with statewide implementation scheduled for 2012–2013.

Pennsylvania use value-added information through their Educational Value-Added Systems (EVAAS), which are based on the TVAAS model in Tennessee. Unlike their Tennessee counterpart, however, value-added reports in the Pennsylvania and North Carolina systems are available only at the level of schools and school cohorts, and are used solely for professional development (McCaffrey and Hamilton 2007). In fact, North Carolina recently adopted new standards for teacher and principal evaluation but omitted a value-added component entirely. As Pennsylvania considers possibilities for a statewide evaluation system that includes student growth measures, similar efforts are under way locally in Pittsburgh through a partnership that includes the school district, the teachers union, and the Bill & Melinda Gates Foundation.[15]

2. Value-Added as a Component in Composite Evaluation Measures

VAM estimates are used along with other measures, such as classroom observations, when they are used at all for evaluating effectiveness. The VAM component in the sites we examined varied from 20 percent to 50 percent of the total rating. For example, the Teacher Effectiveness Measure (TEM) in Memphis assigns a 35 percent weight to student learning growth as measured by TVAAS. The remainder is based on observations of teacher practice (35 percent), stakeholder perceptions (15 percent), and assessments of content and pedagogical knowledge (15 percent).

Both Tennessee and Dallas assign school value-added estimates to principals in measuring their effectiveness. In Tennessee, school value added takes the form of a school-wide TVAAS score. The value-added component of Dallas' Principal Incentive Pay Program (PIPP), the School Effectiveness Index, is simply a version of the CEI that is aggregated to the school level. Principals receive a score on a five-point scale that indicates how far the performance of their school was from the district average. The SEI has a 20 percent weight in the PIPP. The other components include the school's state accountability rating (20 percent) and several measures of school performance (for example, the number of subgroups reaching annual performance targets and the graduation rate) that together account for the remaining 60 percent. The teacher version of the PIPP, the Professional Development and Appraisal System (PDAS), includes four main components: a VAM-based CEI, classroom observations, professional development participation, and teacher attendance.

The weights assigned to components in these models currently vary on an ad hoc basis; research findings do not indicate which weights are optimal. Indeed, differences in local policy preferences could justify differences in weights. At least one recent study asserts that the methodological limitations of VAMs make a 50 percent share unwise (Baker et al. 2010). Analysts are finding, however, that VAM estimates correlate with measures of performance based on professional practice (Grossman et al. 2010; Harris and Sass 2009; Jacob and Lefgren 2008; Rockoff and Speroni 2010).

For example, Jacob and Lefgren (2008) conclude that principals can generally identify the 10 to 20 percent of teachers with the highest and lowest VAM scores. Principals have a harder time distinguishing whether teachers in the middle of the distribution perform above or below average in terms of value added, but analysts encounter the same issues statistically for educators in the middle of the distribution as well (Goldhaber and Hansen 2010; Lipscomb et al. 2010). Relating VAM estimates to other measures of performance appears to provide a fuller description of the characteristics of effective educators, that is—understanding not only who is producing large increases in student achievement but also how their teaching practices may differ (Grossman and

[15] Pittsburgh and Memphis are two of the four sites selected as Intensive Partnership sites by the Bill & Melinda Gates Foundation in 2009. These districts are receiving large grants to recruit, retain, and reward effective teachers.

Loeb 2010; Tyler et al. 2010).[16] Using value-added and professional practice measures together as school districts and states are doing appears also to have the most predictive validity in explaining student achievement gains (Rockoff and Speroni 2010). These analysts study relationships among teacher value-added scores, subjective evaluation scores, and teachers' own value-added scores from earlier years. They find that the relationships are statistically significant when examined separately and, meaningfully, they are very similar when examined together. The findings suggest that each measure might in fact be providing distinct information, at least in their sample. Current findings such as these offer encouragement that, by using both types of measures together, we may improve our ability to identify both effective educators and best practices.

3. Incorporating Staff Across All Grades and Subjects

VAMs such as those in Tennessee, Dallas, and Louisiana apply to many teachers but they do not include everyone. Applicability depends, first of all, on whether a teacher teaches in a tested grade and subject. This restriction affects teachers more than principals because principal VAMs use growth information from students across tested grade levels. A second restriction is that students need a prior score to be included, because VAMs measure performance relative to a baseline. This limitation precludes teachers teaching in the lowest grades in which assessments are administered to students for the first time.

Assessment scores can be used to calculate growth because they are quantifiable and available for students in multiple years. The starting point is usually statewide assessments because they tend to have desirable psychometric properties. Efforts to expand coverage to more grades and subjects then involve increasing the number of assessments that are used. For example, the TEM in Memphis currently applies to 30 percent of teachers but the school district plans to offer additional assessments with the goal of including 65 percent of teachers.

Policymakers and practitioners have adopted strategies for addressing the problem that value added cannot be implemented for all teachers. The first strategy, as used in Memphis, is simply to increase the weight on other evaluation components for teachers lacking an individual VAM score. Another strategy is to apply a school-wide VAM estimate to individual teachers when individual estimates are not possible. Because school districts and states usually offer school performance incentives so that all staff can earn bonuses based on student growth, the second strategy amounts to scaling up the importance of school-wide improvement. Although these approaches might be second best, they are what districts are currently doing and could be effective solutions while methodologies for expanding individual value-added measures are refined.

Despite the lack of full coverage, the implementation of VAMs includes more grades and subjects than analysts' models include in the research literature. For example, TVAAS incorporates reading, language arts, math, science, and social studies in grades 3 to 8 and algebra I, biology, and English II at the high school level. Dallas also applies VAMs directly to foreign language and technology teachers at the middle and high school levels and to teachers in math and reading as early as grade 2. In contrast, nearly all the studies we reviewed examined some variation of math and

[16] Grossman and Loeb (2010) find that higher value-added English language-arts teachers in New York City have a different profile of instructional practices, as measured by the Protocol for Language Arts Teaching Observation (PLATO) and the Classroom Assessment Scoring System (CLASS) than do lower value-added teachers, in particular a great focus on explicit strategy instruction and writing skills. Tyler et al. (2010) find that classroom management and instructional skills, as measured by Cincinnati's Teacher Evaluation System (a Danielson framework), meaningfully predict student achievement growth.

reading teachers in grades 4 to 8. The relatively limited scope of these studies is reflected in analysts' greater interest in scrutinizing aspects of the validity and reliability of value-added methodology itself. That the implementation of VAMs in subjects such as science and social studies, or in early and later grades, might be ahead of the literature is good in many respects but we caution against over-generalizing findings. For example, Mathematica's research in Pittsburgh reveals that value-added measures at the high school level are highly sensitive to rates of student attrition: VAM scores tend to be higher in schools in which more students with low achievement growth profiles drop out (Lipscomb et al. 2010). Despite continual advances in both the research literature and in practice, applying VAMs broadly across grades and subjects remains an important challenge.

CONCLUSION

The studies we reviewed find that value-added measures describe meaningful variation in teacher and principal effectiveness that is not related to students' characteristics or prior achievement trajectories. Analysts have shown that VAMs can distinguish high and low performance from average performance, and that estimates are at least moderately stable over time (with stability increasing as more cohorts of students are added to the estimates). The methods have improved at a rapid pace—as witnessed by growth in both the size of the literature and VAM utilization in school districts and states. But the findings also indicate clearly that continual technical challenges—such as reducing bias and improving reliability, as well as implementation challenges such as applying value added to as many teachers and subjects as possible—remain.

Our review also indicates that the design elements for Pennsylvania's model statewide evaluation system will involve policy decisions and research decisions. Each component, as well as their weights, needs to reflect the state's own policy goals. Even modeling choices such as how many years of teaching data to include and whether to compare teachers within or across schools involve policy tradeoffs for stakeholders to consider. We look forward to working with Team PA and the stakeholder steering committee throughout this pilot study to help inform the development of a valid and reliable state-of-the-art model that serves its intended purpose.

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.

Baker, E.L., P.E. Barton, L. Darling-Hammond, E. Haertel, H.F. Ladd, R.L. Linn, D. Ravitch, R. Rothstein, R.J. Shavelson, and L.A. Shepard. "Problems with the Use of Student Test Scores to Evaluate Teachers." Briefing paper #278. Washington, DC: The Economic Policy Institute, 2010.

Ballou, D. "Value-Added Assessment: Lessons from Tennessee." In *Value Added Models in Education: Theory and Applications*, edited by R. Lissetz (pp. 272–303). Maple Grove, MN: JAM Press, 2005.

Branch, Gregory F., Eric A. Hanushek, and Steven G. Rivkin. "Estimating Principal Effectiveness." Working Paper #32. Washington, DC: The Urban Institute, National Center for Analysis of Longitudinal Data in Education Research, 2009.

Goldhaber, D., and M. Hansen. "Is It Just a Bad Class? Assessing the Stability of Measured Teacher Performance." Working paper 2008-5. Denver, CO: Center for Reinventing Public Education, 2008.

Goldhaber, D. and M. Hansen. "Assessing the Potential of Using Value-Added Estimates of Teacher Job Performance for Making Tenure Decisions." Working paper 31. Denver, CO: Center for Reinventing Public Education, 2010.

Grossman, P., S. Loeb, J. Cohen, K. Hammerness, J. Wyckoff, D. Boyd, and H. Lankford. "Measure for Measure: The Relationship Between Measures of Instructional Practice in Middle School English Language Arts and Teachers' Value-Added Scores." NBER Working Paper #16015. Cambridge, MA: National Bureau of Economic Research, 2010.

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.

Harris, D., and T. Sass. "What Makes for a Good Teacher and Who Can Tell?" Washington, DC: The Urban Institute, National Center for Analysis of Longitudinal Data in Education Research, 2009.

Jacob, Brian A., and Lars 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, Thomas J., and Douglas O. Staiger. "Estimating Teacher Impacts on Student Achievement, An Experimental Evaluation." Working paper no. 14607. Cambridge, MA: National Bureau of Economic Research, 2008.

Koedel, C., and J.R. Betts. "Does Student Sorting Invalidate Value-Added Models of Teacher Effectiveness? An Extended Analysis of the Rothstein Critique." Working paper 902. Columbia, MO: University of Missouri-Columbia, 2009.

Koedel, C., and J.R. Betts. "Value Added to What? How a Ceiling in the Testing Instrument Influences Value-Added Estimation." *Education Finance and Policy,* vol. 5, no. 1, 2010, pp. 54–81.

Lipscomb, S., B. Gill, and K. Booker. "Estimating Teacher and School Effectiveness in Pittsburgh: Value-Added Modeling and Results." Draft report submitted to Pittsburgh Public Schools and the Pittsburgh Federation of Teachers. Cambridge, MA: Mathematica Policy Research, 2010.

Mariano, Louis T., Daniel F. McCaffrey, and J.R. Lockwood. "A Model for Teacher Effects from Longitudinal Data Without Assuming Vertical Scaling." *Journal of Educational and Behavioral Statistics,* vol. 35, no. 3, 2010, pp. 253–279.

McCaffrey, D.F., D. Koretz, J.R. Lockwood, and L.S. Hamilton. "Evaluating Value-Added Models for Teacher Accountability." Santa Monica, CA: RAND Corporation, 2004.

McCaffrey, D.F., and L. Hamilton. "Value-Added Assessment in Practice Lessons from the Pennsylvania Value-Added Assessment System Pilot Project." Santa Monica, CA: RAND Corporation, 2007.

McCaffrey, D.F., T.R. Sass, J.R. Lockwood, and K. Mihaly. "The Inter-Temporal Variability of Teacher Effect Estimates." *Education Finance and Policy,* vol. 4, no. 4, 2009, pp. 572–606.

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

Rockoff, Jonah E., and Cecilia Speroni. "Subjective and Objective Evaluations of Teacher Effectiveness." *American Economic Review: Papers & Proceedings,* vol. 100, May 2010, pp. 261–266.

Rothstein, J. "Student Sorting and Bias in Value-Added Estimation: Selection on Observables and Unobservables." *Education Finance and Policy,* vol. 4, no. 4, 2009, pp. 537–571.

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

Tyler, John H., Eric S. Taylor, Thomas J. Kane, and Amy L. Wooten. "Using Student Performance Data to Identify Effective Classroom Practices." *American Economic Review: Papers & Proceedings,* vol. 100, May 2010, pp. 256–260.

THIS PAGE LEFT BLANK FOR DOUBLE-SIDED PRINTING

APPENDIX A. ARTICLE SUMMARIES

Table A.1. Aaronson, Barrow, and Sander (2007)

Citation	Aaronson, D., Barrow, L., and Sander W. "Teachers and Student Achievement in the Chicago High Schools." *Journal of Labor Economics,* vol. 25, no. 1, 2007, pp. 95–135.
Keywords	Teacher effects, high schools, inter-temporal correlation
Type of VAM	Teacher
Data	
Actual or simulated?	Actual
Location	Chicago
Grades and subjects	Grade 9 in math
School years	1996–1997 to 1998–1999
Outcomes	Math, as measured by the Test of Achievement and Proficiency (TAP)
Control Variables	
Student characteristics	Age, gender, race/ethnicity, meals program, guardian (mother, father, grandparent, and so on), changed high school, repeated ninth grade, math class size
Other controls	Additional specifications control for level and subject matter of math classes, cumulative GPA, class rank, disability status, school outside residential neighborhood, census tract information (median family income, median home value, education level), math class peer average, number of absences, and eighth grade test scores
School effects included?	Some specifications include school effects, but not the primary ones
Model	
Years of baseline test scores	One
Baseline subject controls	Math only (Iowa Test of Basic Skills)
Years of growth data	One
How does study address missing data and/or attrition	Student with missing eighth and ninth grade scores, as well as those in the 1st and 99th percentiles (test score gains), are excluded
Partial year enrollment or team teaching	Restricted to students in self-contained classrooms only; no mention of a dosage approach
Adjustments (such as minimum number of students, shrinkage, or measurement error correction)	15-student minimum for teacher effects to be analyzed; shrinkage estimator
Key Findings	A one standard deviation improvement in math teacher quality raises student scores by one-fifth of average yearly gains. High value-added teachers are particularly important for low ability students. Estimates are stable over time and do not appear to be the result of classroom sorting. Teacher characteristics explain little of the variation in value-added estimates.
Results	
Standard deviation of VAM measures	0.15 (Table 9)
Percentage significant effects	n.a.
Additional years of growth data	n.a.
Inter-temporal correlation	41% of top quartile teachers stay in top quartile the next year; 33% of bottom quartile teachers stay in the bottom quartile the next year
Correlation with other measures	Observed teacher characteristics (for example, gender, race/ethnicity, tenure, advanced degrees, and professional certifications) explain at most 10 percent of variation in teacher quality.
Differences based on choice of outcome	n.a.
Other key findings	The impact of a high quality teacher is largest for African American students and students with low eighth-grade scores, and no different by gender. Value-added measures are not correlated with the fraction of scores that are missing. Teacher assignments are closer to random than sorted, based on prior student achievement in three preceding years.

GPA = grade point Average; VAM = value-added model.

n.a. = not applicable

Table A.2. Ballou (2005)

Citation	Ballou, Dale. "Value-Added Assessment: Lessons from Tennessee." In *Value Added Models in Education: Theory and Applications,* (pp. 272–297), edited by Robert Lissitz. Maple Grove, MN: JAI Press, 2005.
Keywords	Bias, precision, control variables, Tennessee value-added assessment system (TVAAS)
Type of VAM	Teacher
Data	
Actual or simulated?	Actual
Location	Single, moderately large district in Tennessee
Grades and subjects	4–8 in reading and math
School years	1998–1999 to 2000–2001
Outcomes	Reading and math, as measured by the Tennessee Comprehensive Assessment Program (TCAP)
Control Variables	
Student characteristics	Author compares two models: (1) the TVAAS VAM model, which has no controls for student characteristics; and (2) adjusted VAM measures with controls for FRPL status, the percentage of students at the same grade level and school eligible for FRPL, race (white vs. nonwhite), and gender
Other controls	No
School effects included?	No
Model	
Years of baseline test scores	All available years in data
Baseline subject controls	Includes all subjects in data
Years of growth data	Four
How does study address missing data and/or attrition	Students with current year scores and any prior score history are included
Partial year enrollment or team teaching	Students spending at least 150 days assigned to a teacher are included, all others are dropped; dosage used for team teaching
Adjustments (such as minimum number of students, shrinkage, or measurement error correction)	Includes shrinkage adjustment; includes adjustment to TCAP test scores to vertically scale across years and grades
Key Findings	Averaging VAM estimates over three years improves the precision of value-added estimates. Under a three-year model, 58% of grade 7–8 mathematics teachers have significant effects, a gain of 28 percentage points over a one-year model. Within the study district, there is little evidence that omitting student characteristics results in substantial bias under the TVAAS model (which is based on five-year student test-score growth trajectories).
Results	
Standard deviation of VAM measures	n.a.
Percentage significant effects	3% (grade 4–6 grade reading), 8% (grade 7–8 reading), 17% (grade 4–6 math), 30% (grade 7–8 math) using TVAAS single-year VAM estimates and 90% confidence interval
Additional years of growth data	Averaging effects across three years increases the proportion of significant VAM effects to 7%, 11%, 30%, and 58%, respectively
Inter-temporal correlation	Approximately 40% of teachers in the bottom quartile in 1998–1999 remained in the bottom quartile in 1999–2000 (both reading and math, using single-year TVAAS estimates)
Correlation with other measures	n.a.
Differences based on choice of outcome	n.a.
Other key findings	Only one-third of teachers in the study district had enough data in the same subject and grade to generate three-year average VAM estimates. District in study did not exhibit racial or socioeconomic stratification. Highly stratified districts could be more sensitive to the inclusion of controls for student characteristics than the study district.

FRPL = free or reduced-price lunch; VAM = value-added model.

n.a. = not applicable.

Table A.3. Branch, Hanushek, and Rivkin (2009) (findings are preliminary and subject to change)

Citation	Branch, Gregory F., Eric A. Hanushek, and Steven G. Rivkin. "Estimating Principal Effectiveness." Working Paper #32. Washington, DC: The Urban Institute, National Center for Analysis of Longitudinal Data in Education Research, 2009.
Keywords	Principal effectiveness, variation in effectiveness by school poverty level
Type of VAM	Principal
Data	
Actual or simulated?	Actual
Location	Texas
Grades and subjects	3–8 in reading and math
School years	1995–2001
Outcomes	Reading and math, as measures by the Texas Assessment of Academic Skills (TAAS)
Control Variables	
Student characteristics	Gender, race/ethnicity, meals program, special education, ESL, mobility
Other controls	School-by-year proportion low income, classified as special needs, recent immigrants, gender, student mobility (switch to earliest grade offered in different school, switch to other than earliest grade offered in different school); year-by-grade indicators Some specifications include interactions between tenure at a school and the principal-by-school effects
School effects included?	Yes, in some specifications
Model	
Years of baseline test scores	One (expressed in quadratic or cubic form)
Baseline subject controls	Same subject
Years of growth data	One or two
How does study address missing data and/or attrition	n.a.
Partial year enrollment or team teaching	
Adjustments (such as minimum number of students, shrinkage, or measurement error correction)	Shrinkage estimator; observations are weighted to address potential bias in principal value-added estimates arising from having more students with test scores at the lower end of the achievement distribution
Key Findings	Principals, like teachers, generally prefer working in schools with fewer demands as indicated by higher income, higher assessment scores, and smaller minority populations. The tenure of a principal at a school has small but significant impacts on student achievement. The variation in principal effectiveness tends to be the largest in high-poverty schools. Finally, principals who stay in a school tend to be more effective than those who move to other schools, except in the lowest-poverty schools.
Results	
Standard deviation of VAM measures	0.22 (calculated from variance estimate of 0.049)
Percentage significant effects	n.a.
Additional years of growth data	n.a.
Inter-temporal correlation	n.a.
Correlation with other measures	n.a.
Differences based on choice of outcome	n.a.
Other key findings	

ESL = English as a second language; VAM = value-added measure.

n.a. = not applicable.

Table A.4. Goldhaber and Hansen (2008)

Citation	Goldhaber, D., and M. Hansen. "Is It Just a Bad Class? Assessing the Stability of Measured Teacher Performance." Working paper 2008-5. Denver, CO: Center for Reinventing Public Education, 2008.
Keywords	Teacher value added, stability, inter-temporal correlation
Type of VAM	Teacher
Data	
Actual or simulated?	Actual
Location	North Carolina (entire state)
Grades and subjects	Grade 5 in math and reading
School years	1996–1997 to 2005–2006
Outcomes	Math and reading, as measured by North Carolina's annual standardized tests administered in grades 3–8
Control Variables	
Student characteristics	Gender, race and ethnicity, disabilities, free or reduced-price lunch (FRPL) status, parents' education level
Other controls	Teacher characteristics[17] (race and ethnicity, gender, advanced degree, license, NBPTS certification, experience level, mean SAT score of undergraduate institution, certification through approved North Carolina education program, attendance) and school characteristics (school percentage of FRPL-eligible students, school percentage of minority students, class size)
School effects included?	Included only in Appendix A
Model	
Years of baseline test scores	Two
Baseline subject controls	Math and reading
Years of growth data	10
How does study address missing data and/or attrition	Sample includes only students with data for all test scores in grades 3, 4, and 5 in both subjects
Partial year enrollment or team teaching	Sample restricted to self-contained classrooms
Adjustments (e.g. minimum number of students, shrinkage, measurement error correction)	Classroom restrictions limit the size of the class to no fewer than 10 and no more than 29 students
Key Findings	This paper estimates various measures of teacher effectiveness and compares them to assess the extent to which measures of teacher value-added vary over time and across subjects and teaching contexts. The authors find average correlations of 0.3 in reading and 0.5 in math in year-to-year estimates of teacher effectiveness, and a cross-subject correlation that averages near 0.5. The year-to-year variation is greater than what is predicted were random error the only unstable component, implying that teacher job performance does vary over time.
Results	
Standard deviation of VAM measures	0.22 (math) and 0.11 (reading)
Percentage significant effects	n.a.
Additional years of growth data	Additional years of matched teacher-student data increase the precision of estimated teacher effects, but the change in teacher rankings is small.
Inter-temporal correlation	0.32 (reading) and 0.54 (math); simulations suggest upper-bound inter-temporal correlations of 0.52 (reading) and 0.80 (math). The VAM models that include student fixed effects produce teacher effectiveness estimates with considerably lower inter-temporal stability. The year-to-year correlation average is 0.07 for reading and 0.21 for math.
Correlation with other measures	In both subjects, the variation in post-tenure teacher effectiveness is largely unexplained by pre-tenure performance.
Differences based on choice of outcome	n.a.
Other key findings	Introducing a successive year of teaching has a considerable impact on relative rankings—close to 50 percent of teachers change relative rankings by more than one quintile equivalent in math. Math estimates are also more stable over time than estimates in reading. The most-recent past performance estimate is the best predictor of future performance for both subjects. Overall, however, estimates do not support the notion of "stable" performance over time in either subject.

NBPTS = National Board of Professional Teaching Standards; SAT = Scholastic Assessment Test; VAM = value-added model.
n.a. = not applicable.

[17] Teacher characteristic variables are excluded in models with teacher fixed effects.

Table A.5. Goldhaber and Hansen (2010)

Citation	Goldhaber, D., and M. Hansen. "Assessing the Potential of Using Value-Added Estimates of Teacher Job Performance for Making Tenure Decisions." Working paper 2010-31. Denver, CO: Center for Reinventing Public Education, 2010.
Keywords	Teacher effects, tenure decisions
Type of VAM	Teacher
Data	
Actual or simulated?	Actual
Location	North Carolina (entire state)
Grades and subjects	4–5 in math and reading
School years	1995–1996 to 2005–2006
Outcomes	Math and reading, as measured by statewide end-of-year assessment
Control Variables	
Student characteristics	Gender, race/ethnicity, meals program, parent education, grade level
Other controls	No
School effects included?	No
Model	
Years of baseline test scores	One
Baseline subject controls	Both math and reading controls
Years of growth data	One-, two-, and three-year models
How does study address missing data and/or attrition	n.a.
Partial year enrollment or team teaching	Restricted to students in self-contained, nonspecialty classrooms, no dosage approach
Adjustments (such as minimum number of students, shrinkage, or measurement error correction)	10 student minimum for teacher effects to be analyzed; shrinkage estimator used
Key Findings	The variation in teacher value-added changes little over the course of teachers' careers. VAM estimates in prior years, even those with a multiyear lag, are good predictors of current year value-added. VAMs provide useful information for teacher evaluation.
Results	
Standard deviation of VAM measures	0.22 (math) and 0.10 (reading) for a one-year model
Percentage significant effects	n.a.
Additional years of growth data	Three-year effects have higher predictive power five years later than one-year effects one year later.
Inter-temporal correlation	0.53 (math) for one-year model. Inter-temporal correlations do not fall to zero even after nine years.
Correlation with other measures	Prior-year VAM scores, even those in other subjects, predict later value-added. Adding controls for observable teacher characteristics (for example, experience or credential) does not change these relationships.
Differences based on choice of outcome	n.a.
Other key findings	

Note: Sections of this paper are available in published form: Goldhaber, D., and M. Hansen. "Using Performance on the Job to Inform Teacher Tenure Decisions." *American Economic Review: Papers & Proceedings,* vol. 100, 2010, pp. 250–255.

VAM = value-added model.

n.a. = not applicable.

Table A.6. Grossman et al. (2010)

Citation	Grossman, P., S. Loeb, J. Cohen, K. Hammerness, J. Wyckoff, D. Boyd, and H. Lankford. "Measure for Measure: The Relationship between Measures of Instructional Practice in Middle School English Language Arts and Teachers' Value-Added Scores." NBER Working Paper #16015. Cambridge, MA: NBER, 2010.
Keywords	Teacher value added, instructional practices, classroom practice measures
Type of VAM	Teacher [a]
Data	
Actual or simulated?	Actual
Location	New York City
Grades and subjects	Grades 6,7, and 8; English Language Arts (ELA)
School years	2006–2007
Outcomes	ELA, as measured by a standardized assessment (no further details). The first functional form expresses the outcome in gains form. The second specification uses the ELA score as the outcome and controls for prior scores.
Control Variables	
Student characteristics	Two separate VAMs using different controls: 1. Student fixed effects and student time-varying characteristics (for example, student changes schools). This strategy identifies VA by comparing teachers who teach the same students, usually in different years. 2. Gender, race, eligibility for FRPL, prior-year test scores in math and ELA, and ELL status, among other factors.
Other controls	1. School characteristics, classroom characteristics, and indicators for year and grade; 2. Classroom characteristics (aggregates of the student controls, standard deviation of prior scores); school variables (enrollment, percentage of black and Hispanic students, percentage ELL, school average expenditures per pupil); indicators for year and grade
School effects included?	No
Model	
Years of baseline test scores	Not reported
Baseline subject controls	Math and ELA (for 2nd VA specification).
Years of growth data	Not reported
How does study address missing data and/or attrition	Not reported
Partial year enrollment or team teaching	Not reported
Adjustments (such as minimum number of students, shrinkage, or measurement error corr.)	Empirical Bayes shrinkage estimator
Key Findings	High VA teachers have a different profile of instructional practices—as measured by 16 practice-based elements from the Protocol for Language Arts Teaching Observation (PLATO) and the Classroom Assessment Scoring System (CLASS)—than low VA teachers. Teachers in the 4th (top) VA quartile score higher than 2nd-quartile teachers on all 16 elements. Elements include clarity of lesson purpose, intellectual challenge in assignments, quality feedback, and presence of an explicit strategy instruction. Differences are statistically significant for explicit strategy instruction and approach statistical significance for intellectual challenge and guided practice.
Results	
Standard deviation of VAM measures	Not reported
Additional years of growth data	n.a.
Correlation with other measures	Teachers in the top (4th) VA quartile score higher than those in the 2nd quartile on all elements studied.
Differences based on choice of outcome	n.a.
Other key findings	The study finds that high VA teachers, relative to low value-added teachers, report focusing more on writing and research skills than on reading.

[a] This paper evaluates what classroom practices differentiate teachers with high VA scores from teachers with low VA scores. The researchers first estimate teacher VAMs for all New York City teachers of middle school ELA. Next, they select 24 3rd- to 5th-year teachers scoring in the fourth and second quartiles for the evaluation of instructional practices.

ELL = English language learner; FRPL = free or reduced-price lunch; VA = value-added; VAM = value-added model; n.a. = not applicable.

Table A.7. Hanushek and Rivkin (2008)

Citation	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.
Keywords	Teacher quality variation, school sorting, teacher mobility, school characteristics
Type of VAM	Teacher
Data	
Actual or simulated?	Actual
Location	One large urban district in Texas
Grades and subjects	4–8 in math
School years	1995–1996 to 2000–2002
Outcomes	Math, as measured by the Texas Assessment of Academic Skills (TAAS)
Control Variables	
Student characteristics	Model includes unspecified nonschool factors
Other controls	Model includes unspecified peer and school factors
School effects included?	Model is estimated both with and without school fixed effects.
Model	
Years of baseline test scores	One
Baseline subject controls	Math only
Years of growth data	Five years
How does study address missing data and/or attrition	Not reported
Partial year enrollment or team teaching	Not reported
Adjustments (such as minimum number of students, shrinkage, or measurement error correction)	Test measurement error and nonpersistent year-to-year fluctuations in teacher effectiveness are addressed by using multiple years of data. The study examines the potential sorting of students to classrooms based on prior achievement by comparing the variation in teacher effectiveness separately for "sorting" schools and "nonsorting" schools and by conducting falsification tests.
Key Findings	The distribution of teacher effectiveness ranges from 0.13 to 0.20 of a standard deviation, implying that moving from an average teacher to a teacher at the 84th percentile of the quality distribution would move a student from the 50th percentile to the 55th (58th) percentile. The range of estimates is based on whether school effects are included (school effects lower variability). There is some evidence of classroom sorting but little evidence that it biases VA measures. Teachers who exit teaching appear significantly less effective than those who stay.
Results	
Standard deviation of VAM measures	0.13–0.20
Percentage significant effects	Not reported
Additional years of growth data	Reported results are VAM averages across 5 years
Inter-temporal correlation	Approximately 0.4, though it varies somewhat with the sample used for the estimation
Correlation with other measures	n.a.
Differences based on choice of outcome	n.a.
Other key findings	

Note: Additional information can be found at www.caldercenter.org/upload/Eric_Hanushek_presentation.pdf.

VA = value added; VAM = value-added model.

n.a. = not applicable.

Table A.8. Harris and Sass (2009)

Citation	Harris, D., and T. Sass. "What Makes for a Good Teacher and Who Can Tell?" Washington, DC: The Urban Institute, National Center for Analysis of Longitudinal Data in Education Research, 2009.
Keywords	Teacher value added, principal evaluations, teacher characteristics
Type of VAM	Teacher
Data	
Actual or simulated?	Actual
Location	A midsize school district in Florida
Grades and subjects	Grades 2 through 10 in Math and Reading
School years	2000–2001 through 2007–2008
Outcomes	Math and reading gains, as measured by the Stanford Achievement Test
Control Variables	
Student characteristics	Student effects and time-varying student characteristics (such as mobility)
Other controls	Peer characteristics (includes both exogenous peer characteristics and the number of peers or class size)
School effects included?	Yes
Model	
Years of baseline test scores	One
Baseline subject controls	None; prior scores are on the left-hand side because the outcome is in the form of test score gains
Years of growth data	Varies: - 1-year estimates for 2005–2006 - 2-year estimates for 2004–2005 through 2005–2006, and 2006–2007 through 2007–2008 - 6-year estimates for 2000–2001 through 2005–2006
How does study address missing data and/or attrition	Not reported
Partial year enrollment or team teaching	No dosage measure
Adjustments (such as minimum number of students, shrinkage, or measurement error corr.)	- Sample is limited to teachers who taught at least 5 students with achievement gain data. - Empirical Bayes shrinkage technique is used. - Student effects control for classroom sorting on time-invariant characteristics. - Falsification tests are used to detect evidence of bias due to classroom sorting based on time-varying factors (for example, high-achieving students assigned nonrandomly to certain teachers). These tests are based on the idea that future teachers cannot have causal effects on current achievement gains. The study finds no evidence that their data are subject to dynamic sorting bias.
Key Findings	Teacher VA and principals' subjective ratings are positively correlated. Principals' evaluations are better predictors of a teacher's VA than traditional approaches to teacher compensation focused on experience and formal education. Teachers' subject knowledge, teaching skill, and intelligence are most closely associated with both the overall subjective teacher rating and the teacher value added. Although prior teacher VA predicts future teacher VA, the principals' subjective ratings can provide additional information and increase predictive power.
Results	
Standard deviation of VAM measures	n.a.
Percentage significant effects	n.a.
Additional years of growth data	Prior VA is a stronger predictor when multiple student cohorts are included in the prior VA measure than in a one-year model.
Inter-temporal correlation	Study reports bivariate regression coefficients between current and prior VA Correlations are statistically significant if based on two or more years of growth data.
Correlation with other measures	Teacher VA and principals' subjective ratings are positively correlated.
Differences based on choice of outcome	n.a.
Other key findings	Principal ratings outperform past VA when VA measures use a single year of data, as they do for new teachers. Prior VA does a better job at predicting VA than principal ratings when multiple years of data are used.

VA = value added; VAM = value-added model; n.a. = not applicable.

Table A.9. Jacob and Lefgren (2008)

Citation	Jacob, Brian A., and Lars 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.
Keywords	Teacher value added, principal evaluations, teacher characteristics
Type of VAM	Teacher
Data	
Actual or simulated?	Actual
Location	Midsize school district in the western United States
Grades and subjects	Grades 2–6
School years	2003–2004 to 2004–2005
Outcomes	Math and reading scores, as measured by district core exams
Control Variables	
Student characteristics	Age, race, gender, FRPL eligibility, special education placement, limited English proficiency status, and grade level
Other controls	Class size, class average student characteristics and prior achievement, school year
School effects included?	Yes
Model	
Years of baseline test scores	One
Baseline subject controls	Both math and reading
Years of growth data	Six (1997–1998 to 2002–2003)
How does study address missing data and/or attrition	Not reported
Partial year enrollment or team teaching	Not reported
Adjustments (such as minimum number of students, shrinkage, or measurement error correction)	Shrinkage estimator
Key Findings	Principals can generally identify teachers who produce the largest and smallest standardized achievement gains in math and reading (top and bottom 10 to 20 percent) but have far less ability to distinguish between teachers in the middle of this distribution (middle 60 to 80 percent). Previous value added is a better predictor of value added than are contemporaneous principal evaluations.
Results	
Standard deviation of VAM measures	0.12 in reading and 0.26 in math
Percentage significant effects	n.a.
Additional years of growth data	n.a.
Inter-temporal correlation	n.a.
Correlation with other measures	Correlation between a principal's evaluation of how effective a teacher is at raising student achievement and the teacher's VA is 0.29 (reading) and 0.32 (math). Correlation does not appear to vary systematically with experience (after the first year), the duration the principal has known the teacher, compensation, or grade taught.
Differences based on choice of outcome	The authors use several transformations of students' test scores as their outcome variables: the percentage correct score, including a student's percentile rank within his year and grade, the square of the percentage correct, and the natural logarithm of the percentage correct. Main results are robust to these alternative VA measures.
Other key findings	Principal assessments are a much better predictor of future student achievement than are traditional measures of teacher compensation, such as education and experience.

FRPL = free or reduced-price lunch; VA = value added; VAM = value-added model.

n.a. = not applicable.

Table A.10. Kane and Straiger (2008)

Citation	Kane, T., and D. O. Staiger. "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation. NBER working paper #14607. Cambridge, MA: National Bureau of Economic Research, 2008.
Keywords	Teacher effects, random assignment of teachers to classrooms, fade out
Type of VAM	Teacher
Data	
Actual or simulated?	Actual
Location	Los Angeles (primary location); comparative non-experimental analyses in New York City and Boston
Grades and subjects	Grades 2 to 5 in math and reading
School years	2000–2007 (Los Angeles); 2000–2006 (New York City), 2006–2007 (Boston)
Outcomes	Math and reading, as measured by the Stanford 9 (1999–2002), the California Achievement Test (2003), and the California Standards Test (2004–2007)
Control Variables	
Student characteristics	Race/ethnicity, grade repetition, Title I status, meals program, homeless, migrant, gifted, disability, English language development, grade
Other controls	Classroom-level means of student variables
School effects included?	Yes, in some specifications
Model	
Years of baseline test scores	One
Baseline subject controls	Math and reading
Years of growth data	One
How does study address missing data and/or attrition	Student with missing scores are excluded, as well as students in classrooms in which more than 20 percent are identified as special education students.
Partial year enrollment or team teaching	No mention of a dosage approach
Adjustments (such as minimum number of students, shrinkage, or measurement error correction)	10 student minimum for teacher effects to be analyzed; teachers must have at least three years of experience; shrinkage estimator
Key Findings	Cannot reject that non-experimental teacher effects (that is, those estimated without random assignment of teachers to classrooms) are unbiased predictors of student achievement gains later under random assignment. Conditioning on prior scores appears to remove any bias due to nonrandom assignment. Teacher effects fade out at a rate of roughly 50 percent per year.
Results	
Standard deviation of VAM measures	0.16 to 0.19 (math); 0.13 to 0.16 (reading) for Los Angeles, New York City, and Boston; standard deviations in Los Angeles are nearly as large by adding school fixed effects
Percentage significant effects	n.a.
Additional years of growth data	n.a.
Inter-temporal correlation	n.a.
Correlation with other measures	n.a.
Differences based on choice of outcome	n.a.
Other key findings	Findings from Los Angeles, New York City, and Boston suggest a similar degree of tracking of students into classrooms based on their expected baseline achievement. Despite this, there is almost no correlation between students' baseline achievement and the effectiveness of the teacher to whom they were assigned.

VAM = value-added model.

n.a. = not applicable.

Table A.11. Koedel and Betts (Forthcoming)

Citation	Koedel, C., and J. R. Betts. "Does Student Sorting Invalidate Value-Added Models of Teacher Effectiveness? An Extended Analysis of the Rothstein Critique." *Education Finance and Policy,* forthcoming.
Keywords	Bias in value-added estimates, nonrandom sorting of students
Type of VAM	Teacher
Data	
Actual or simulated?	Actual
Location	San Diego Unified School District
Grades and subjects	Grade 4 in math
School years	1998–1999 to 2001–2002
Outcomes	Math, as measured by Stanford 9 (vertically scaled)
Control Variables	
Student characteristics	Gender, race, ELL status, change from ELL to English-proficient status, expected and unexpected school changer, parental education, designated as advance student, percentage of school year absent
Other controls	Model 1: year-specific common intercept across students Model 2: Model 1 + school-level covariates (percentage of student body by race, ELL status, FRPL status and school-changer status) + classroom-level covariates (classroom-level peer performance in baseline year, class size) and school fixed effects Model 3: Model 2 + student fixed effects Model 4: Instrument lagged test score gain with second-lagged test score level
School effects included?	Yes (models 2,3, and 4)
Model	
Years of baseline test scores	One or two
Baseline subject controls	Math
Years of growth data	One or two
How does study address missing data and/or attrition	Includes only students with 4th-grade and lagged test scores (second-lagged test scores for model 2)
Partial year enrollment or team teaching	
Adjustments (such as minimum number of students, shrinkage, or measurement error correction)	Restricted sample to teachers with at least 20 students across data panel and students taught by these teachers; included students who repeated 4th grade Excluded students who switched schools (following Rothstein)
Key Findings	A sufficiently complex VAM that evaluates teachers using multiple years of teaching data reduces the sorting-bias problem to statistical insignificance. Although data for the first couple of years for novice teachers might be insufficient to measure teacher quality effectively, value added continues to provide useful information for other teachers.
Results	
Standard deviation of VAM measures	0.18–0.24 (4th-grade math teachers) 0–0.15 (5th-grade math teachers)
Percentage significant effects	n.a.
Additional years of growth data	Reduces transitory sorting bias significantly
Inter-temporal correlation	n.a.
Correlation with other measures	n.a.
Differences based on choice of outcome	n.a.
Other key findings	n.a.

ELL = English language learner; FRPL = free or reduced-price lunch; VAM = value-added model.

n.a. = not applicable.

Table A.12. Koedel and Betts (2010)

Citation	Koedel, C., and J. R. Betts. "Value Added to What? How a Ceiling in the Testing Instrument Influences Value-Added Estimation." *Education Finance and Policy*, vol. 5, no. 1, 2010, pp. 54–81.
Keywords	Value added, teacher, education, testing instrument, ceiling effects
Type of VAM	Teacher
Data	
Actual or simulated?	Actual data provide the no-ceiling baseline VA estimates. Test score ceilings are simulated on the actual scores and their effects are evaluated by comparing VA estimates simulated under the ceilings with the baseline VA estimates.
Location	San Diego
Grades and subjects	4th-grade math
School years	1998–1999 through 2001–2002
Outcomes	Math, as measured by the Stanford 9 test. Test-score ceiling conditions are simulated on the Stanford 9 scores based on the skewness of score distributions from the math sections of the Texas Assessment of Academic Skills (TAAS) and the Florida Comprehensive Assessment Test (FCAT).
Control Variables	
Student characteristics	Controls vary by specification: (1) and (2): race, gender, ELL status, change from ELL to English proficient, expected and unexpected school changer, designated as advanced student, percentage of school year absent, and parental education; (3): student effects and time-varying student characteristics
Other controls	(1): none; (2) and (3): classroom-level peer performance in prior year, class size, percentage of student body by race, ELL status, meal program status, school changers
School effects included?	(1) no school effects; (2) and (3) include school effects
Model	
Years of baseline test scores	(1) and (2): one; (3): two
Baseline subject controls	Math only
Years of growth data	One
How does study address missing data and/or attrition	Students without test score records in two contiguous grades (three grades for specification 3) are excluded.
Partial year enrollment or team teaching	Not reported
Adjustments (such as minimum number of students, shrinkage, or measurement error corr.)	VA is estimated for teachers with at least 20 students. Student sample is restricted to students taught by those teachers. Adjusted estimates account for estimation error in teacher effects.
Key Findings	Over a wide range of test score ceiling severity, teachers' VA estimates are negligibly influenced by ceiling effects. For example, a ceiling capping maximum scores at the 75th percentile is fairly inconsequential, with a high correlation (0.92–0.94) between baseline VA estimates and simulated VA estimates. As ceiling conditions approach the severity of those found in *minimum-competency* testing, VA results are significantly altered. For example, a ceiling capping maximum scores at the 33rd percentile leads to lower correlations between baseline and simulated VA estimates (0.72–0.77).
Results	
Standard deviation of VAM measures	VA estimates are reported in terms of adjusted effect sizes of teacher quality, in which the adjustment accounts for estimation error in the individual teacher effect estimates. Adjusted effect sizes range from 0.22 to 0.25 standard deviations, regardless of test score ceilings.
Percentage significant effects	n.a.
Additional years of growth data	n.a.
Inter-temporal correlation	n.a.
Correlation with other measures	n.a.
Differences based on choice of outcome	n.a.
Other key findings	1. Minimum-competency ceiling effects significantly alter teachers' VA *rankings*. For example, only 49–56 percent of teachers ranked in the top 20 percent of the baseline VA distribution are also ranked in that quintile after maximum ceiling effects are imposed. 2. The estimation error share of the variance of teacher effects increases as ceiling severity increases.

ELL = English language learner; VA = value added; VAM = value-added model; n.a. = not applicable.

Table A.13. Lipscomb, Gill, and Booker (2010)

Citation	Lipscomb, S., B. Gill, and K. Booker, K. "Estimating Teacher and School Effectiveness in Pittsburgh: Value-Added Modeling and Results." Draft report submitted to Pittsburgh Public Schools and the Pittsburgh Federation of Teachers. Cambridge, MA: Mathematica Policy Research, 2010.
Keywords	Distribution of teacher and school effects, precision, inter-temporal correlation, school effects, anchoring to state distribution, high school attrition
Type of VAM	Teacher, school
Data	
Actual or simulated?	Actual
Location	Pittsburgh
Grades and subjects	4–8 in math and reading (teacher and school analyses); 9 and 11 (school only)
School years	2006–2007 to 2008–2009
Outcomes	Math and reading (grades 4–8, 11), as measured by the Pennsylvania System of School Assessment (PSSA); Algebra I and reading (grade 9), as measured by curriculum-based assessments
Control Variables	
Student characteristics	Gender, race/ethnicity, meals program, ELL status, disability, gifted, grade repeater, prior absences and suspensions, mobility, grade level
Other controls	Class size, class average prior PSSA scores, and class averages for the above student characteristics
School effects included?	Not in main model
Model	
Years of baseline test scores	One; supplementary analyses include two
Baseline subject controls	Both math and reading controls
Years of growth data	One-, two-, and three-year models
How does study address missing data and/or attrition	Students are dropped if missing prior- or current-year assessment scores
Partial year enrollment or team teaching	Dosage model
Adjustments (such as minimum number of students, shrinkage, or measurement error corr.)	10 student minimum for teacher effects to be analyzed; shrinkage estimator
Key Findings	Variation in teacher and school effectiveness is consistent with prior studies. Precision improves with multiple years of student cohort data. There is more variation in effectiveness within schools than across them. Inter-temporal correlations are higher than in related metrics used in Pittsburgh. Attrition is a challenge for VAMs at the high school level.
Results	
Standard deviation of VAM measures	0.18 (math) and 0.18 (reading) for a one-year teacher model. 0.14 (math) and 0.11 (reading) for a one-year school model
Percentage significant effects	41% (math) and 29% (reading) for a one-year teacher model and 90% confidence interval; 54% (math) and 54% (reading) for a one-year school model
Additional years of growth data	63% (math) and 42% (reading) for a three-year teacher model; 69% (math) and 65% (reading) for a one-year school model
Inter-temporal correlation	0.58 (math) and 0.33 (reading) for teachers; 42% to 56% of top-quartile teachers stay in top quartile the next year; 39% to 40% of bottom-quartile teachers stay in bottom quartile the next year; 0.65 (math) and 0.48 (reading) for schools; 38% to 54% of top-quartile schools stay in top quartile the next year; 50% to 54% of bottom-quartile schools stay in bottom quartile the next year.
Correlation with other measures	School VAMs correlate with PULSE (the district's principal bonus allocation plan) and the Pennsylvania Value-Added Assessment System (PVAAS), particularly in math.
Differences based on choice of outcome	Models based on PSSA scores have higher predictive ability than CBA in high school analyses.
Other key findings	Attrition at the high school level leads to biased value-added estimates because schools with high sample attrition have high VAM scores (as calculated based on the students who remain enrolled).

CBA = curriculum-based assessments; ELL = English language learner; VAM = value-added model.

Table A.14. Mariano, McCaffrey, and Lockwood (2010)

Citation	Mariano, Louis T., Daniel F. McCaffrey, and J. R. Lockwood. " A Model for Teacher Effects From Longitudinal Data Without Assuming Vertical Scaling." *Journal of Educational and Behavioral Statistics,* vol. 35, no. (3, 2010, pp. 253–279.
Keywords	teacher effects, value-added models, vertical scaling, Bayesian methods
Type of VAM	Teacher
Data	
Actual or simulated?	Actual
Location	Large urban school district (location not reported)
Grades and subjects	Mathematics in grades 1 to 5 (for 1 student cohort)
School years	1997–1998 through 2001–2002
Outcomes	Vertically scaled annual mathematics scores from a national commercial assessment (no further detail)
Control Variables	
Student characteristics	None
Other controls	Year indicators to control for overall mean scores and indicators for current and prior teachers
School effects included?	The theoretical section mentions an "extended" version of the "generalized persistence" model that includes student characteristics and school effects. However, the empirical model uses the basic version without such controls.
Model	
Years of baseline test scores	None (all test scores are used jointly in the estimation of the "generalized persistence" model)
Baseline subject controls	None
Years of growth data	One, because the study follows a single cohort of students through time
How does study address missing data and/or attrition	- Missing test score data are accommodated by data augmentation, assuming that missing scores are missing at random - Missing student-teacher links are set to a dummy teacher with zero effect.
Partial year enrollment or team teaching	Empirical example assumes that each student had only one teacher each year, but the model can be modified to allow for multiple teachers and teachers with part-year contributions.
Adjustments (such as minimum number of students, shrinkage, or measurement error corr.)	Not reported
Key Findings	This study relaxes the assumption that teacher effects persist undiminished in future years or that they diminish but remain perfectly correlated. This assumption might be inconsistent with assessment data that are not vertically scaled or in which the mix of topics changes as students advance in school. The proximal-year effects have much larger variation than future-year effects, suggesting that complete persistence of teacher effects across future years is not supported by the data. The assumption of perfect correlation between proximal and future effects is also not entirely consistent with the data.
Results	
Standard deviation of VAM measures	Not reported
Percentage significant effects	n.a.
Additional years of growth data	n.a.
Inter-temporal correlation	Proximal-year effects are estimated to have correlations of about 0.5 to 0.6 with the future-year effects (though the latter are estimated from future grades in which the students taught by the evaluated teacher in the proximal year are now taught by different teachers). Future-year effects are estimated to have correlations of about 0.9 or higher among themselves. This suggests that while the effect a teacher has on his or her students in the proximal year is different from future effects, future effects are very similar to one another.
Correlation with other measures	n.a.
Differences based on choice of outcome	n.a.
Other key findings	The proximal effect estimates from the generalized persistence model in this study were highly correlated with those from simpler variable persistence models, suggesting that the latter models can be used for proximal year inferences.

VAM = value-added model; n.a. = not applicable.

Table A.15. McCaffrey and Hamilton (2007)

Citation	McCaffrey, D. F., L. and Hamilton, L. "Value-Added Assessment in Practice: Lessons from the Pennsylvania Value-Added Assessment System Pilot Project. Santa Monica, CA: RAND Corporation, 2007.
Keywords	PVAAS, student achievement growth
Type of VAM	School
Data	
Actual or simulated?	Actual
Location	93 districts in Pennsylvania (47 PVAAS pilot districts and 46 matched comparison districts)
Grades and subjects	5 and 8 in math and reading
School years	2005–2006
Outcomes	Math and reading (grades 5 and 8), as measured by the Pennsylvania System of School Assessment (PSSA)
Control Variables	
Student characteristics	None
Other controls	None
School effects included?	No
Model	
Years of baseline test scores	Three total prior scores either from prior years or across subjects
Baseline subject controls	See above
Years of growth data	One
How does study address missing data and/or attrition	Uses all available data to estimate model coefficients using incomplete data methods
Partial year enrollment or team teaching	n.a.
Adjustments (such as minimum number of students, shrinkage, or measurement error correction)	None
Key Findings	Participation in PVAAS did not raise or lower student achievement in pilot districts. PVAAS information was not being used by administrators, principals, and teachers in significant ways. Lack of stakes attached to PVAAS might contribute to low usage. Awareness and engaged use of PVAAS might improve in subsequent years.
Results	
Standard deviation of VAM measures	n.a.
Percentage significant effects	n.a.
Additional years of growth data	n.a.
Inter-temporal correlation	n.a.
Correlation with other measures	n.a.
Differences based on choice of outcome	n.a.
Other key findings	Tennessee and Dallas provide individual reports to teachers. North Carolina, Ohio, and Pennsylvania do not.

PVAAS = Pennsylvania Value-Added Assessment System; VAM = value-added model.

n.a. = not applicable.

Table A.16. McCaffrey, Lockwood, and Mihaly (2009)

Citation	McCaffrey, D. F., T. R. Sass, J. R. Lockwood, and K. Mihaly. "The Inter-Temporal Variability of Teacher Effect Estimates." *Education Finance and Policy,* vol. 4, no. 4, 2009, pp. 572–606.
Keywords	Inter-temporal correlation, teacher characteristics, averaging estimates from two years
Type of VAM	Teacher
Data	
Actual or simulated?	Actual
Location	5 large Florida districts (Dade, Duval, Hillsborough, Orange, and Palm Beach)
Grades and subjects	4–8 in math
School years	2000–2001 to 2004–2005
Outcomes	Math, as measured by the Florida Comprehensive Achievement Test - Norm Referenced Test (FCAT-NRT) [primary outcome] and FCAT Sunshine State Standards Test (FCAT-SSS)
Control Variables	
Student characteristics	Gender, race/ethnicity, meals program, gifted program, LEP, disability, mobility
Other controls	Class average shares by gender, Black, changed schools, age, class size
School effects included?	No
Model	
Years of baseline test scores	One
Baseline subject controls	Math only
Years of growth data	One or two
How does study address missing data and/or attrition	Students with incomplete information are excluded
Partial year enrollment or team teaching	Restricted to students in self-contained classrooms only; no mention of a dosage approach
Adjustments (such as minimum number of students, shrinkage, or measurement error correction)	15 student minimum for teacher effects to be analyzed; shrinkage estimator
Key Findings	The year-to-year correlation of teacher value added ranges from 0.2–0.5 in elementary grades and from 0.3–0.7 for middle school grades. Teacher effects are moderately stable, with correlations in the range found in other occupations. Stability improves up to 40 to 60 percent by averaging teacher effects across two years.
Results	
Standard deviation of VAM measures	n.a.
Percentage significant effects	n.a.
Additional years of growth data	Stability improves up to 40 to 60 percent by averaging teacher effects across two years
Inter-temporal correlation	0.2–0.5 (elementary), 0.3–0.7 (middle); one-third of top- (bottom-) quintile teachers stay in top (bottom) quintile the next year; correlations improve from 40 to 60% when a two-year average of teacher effects is used
Correlation with other measures	Observed teacher characteristics (experience, advanced degrees, professional development) explain little of inter-temporal variation unrelated to sampling errors
Differences based on choice of outcome	Inter-temporal correlations do differ if the high stakes assessment is used, but no clear pattern emerges across districts
Other key findings	30 to 60 percent of variation in measured teacher value added is due to sampling error; persistent teacher effects make up 50 to 70 percent of remainder; other time-varying factors account for rest. Inter-temporal correlations are within range found in other occupations.

LEP = limited English proficiency; VAM = value-added model.

n.a. = not applicable.

Table A.17. Potamites, Booker, Chaplin, and Isenberg (2009)

Citation	Potamites, L., K. Booker, D. Chaplin, and E. Isenberg. Measuring School and Teacher Effectiveness in the EPIC Charter School Consortium – Year 2. Washington, DC: Mathematica Policy Research, 2009.
Keywords	EPIC charter school consortium, shrinkage estimator, test score standardization, partial dosage
Type of VAM	School, teacher
Data	
Actual or simulated?	Actual
Location	17 states (CA, CO, FL, GA, HI, IL, IN, LA, MA, MI, MN, MO, NM, NY, OH, PA, and TX,) and the District of Columbia
Grades and subjects	Elementary, middle, and high school in reading and math
School years	2006–2007 and 2007–2008
Outcomes	State test scores (standardized across states, grades, and years using statewide means and SDs, and state and national NAEP means and SDs)
Control Variables	
Student characteristics	Gender, race/ethnicity, FRPL, LEP, special education, first year at new school, skipping/failing grade since most recenttest
Other controls	Grade level, subject, year, interactions, flag for educated guess about testing scale used in FL
School effects included?	No
Model	
Years of baseline test scores	One
Baseline subject controls	Other subject is used as instrumental variable for same-subject test score (for example, math as instrument for reading when outcome is reading)
Years of growth data	One and two
How does study address missing data and/or attrition	Missing data imputed for demographics only; no imputation for missing test scores
Partial year enrollment or team teaching	School dosage variable constructed; equal to the percentage of the year student spent at school; teacher dosage variable constructed based on days enrolled in teacher's classroom; equal to the proportion of the year spent with that teacher (set to zero if fewer than two weeks and set to one if all but two weeks or fewer spent in teacher's classroom)
Adjustments (such as minimum number of students, shrinkage, or measurement error correction)	15 students minimum for estimating teacher and school effects. Measurement error controlled for by using a 2SLS model with the student's prior test score in the other subject as an instrument for the prior same-subject test score. Shrinkage estimator used.
Key Findings	Averaging VA estimates across two years decreases the fraction of their variance that is due to noise and improves reliability. The mean standard error of VA estimates under a two-year model falls relative to mean standard error in a one-year model (that is, from 0.081 to 0.51 for schools and from 0.160 to 0.132 for teachers). At each grade level, the top-ranked teacher scores higher than the lowest-ranked teacher, but teachers ranked near each other generally are not statistically distinguishable.
Results	
Standard deviation of VAM measures	Overall teacher VAM: SD = 0.261 for full one-year model and SD = 0.235 for full two-year model. Overall school VAM: SD = 0.187 for full one-year model and SD = 0.191 for full two-year model. Variation is largest at the high school level.
Percentage significant effects	n.a.
Additional years of growth data	Mean SEs for the school estimates decrease from 0.081 for the one-year model to 0.051 for the two-year model. Mean SEs for the teacher estimates decrease from 0.160 for the one-year model to 0.132 for the two-year model.
Inter-temporal correlation	n.a.
Correlation with other measures	The one- and two-year estimates for schools were correlated at 0.94.
Differences based on choice of outcome	n.a.
Other key findings	n.a.

2SLS = two-stages least square; EPIC = Effective Practices Incentive Community; FRPL = free or reduced-price lunch; LEP = limited English proficiency; NAEP = National Assessment of Educational Progress; SD = standard deviation; SE = standard error; VA = value added; VAM = value-added model; n.a. = not applicable.

Table A.18. Rockoff and Speroni (2010)

Citation	Rockoff, Jonah E., and Cecilia Speroni. "Subjective and Objective Evaluations of Teacher Effectiveness." Am*erican Economic Review: Papers & Proceedings,* vol. 100, May 2010, pp. 261–266.
Keywords	Teacher effectiveness, subjective evaluations, objective evaluations
Type of VAM	Teacher [a]
Data	
Actual or simulated?	Actual
Location	New York City
Grades and subjects	Grades 3 to 8; math and English
School years	2003–2004 through 2007–2008
Outcomes	Math and English standardized test scores (no further detail provided)
Control Variables	
Student characteristics	Gender, race, prior suspensions and absences, and indicators for English Language Learner, Special Education, grade retention, and free or reduced-price lunch status. These controls are also interacted with grade level.
Other controls	Teacher experience; classroom and school-year averages of student characteristics; class size; year-grade and zip code fixed effects
School effects included?	No
Model	
Years of baseline test scores	One
Baseline subject controls	Math and English (in cubic polynomial form); not reported whether both are included in the VA estimation for each subject
Years of growth data	One year (each teacher's first year of teaching)
How does study address missing data and/or attrition	Not reported
Partial year enrollment or team teaching	Not reported
Adjustments (such as minimum number of students, shrinkage, or measurement error corr.)	Not reported
Key Findings	This study examines the relationship between VAM-based evaluations and subjective evaluations. Subjective measures include mentor and teaching fellow interviewer evaluations. When examined separately, higher VAM-based evaluations of first-year performance and higher subjective evaluation scores are each related to higher student achievement growth in a teacher's second year. When the two types of evaluations are examined together, "their coefficients are only slightly attenuated—each evaluation contains information distinct from the other."
Results	
Standard deviation of VAM measures	These are VA estimates' coefficients in predicting second-year student test scores, **not** the sizes of the VA estimates themselves (those are not reported):[a] All teachers: 0.088 (math), 0.02 (English) – significant at 5% NYC Teaching Fellows: 0.095 (math) – significant at 5%; English not reported Mentored teachers: 0.085 (math) – significant at 5%; English not reported
Percentage significant effects	n.a.
Additional years of growth data	n.a.
Inter-temporal correlation	n.a.
Correlation with other measures	Relationships are evaluated between New York City teaching fellow/mentors' subjective evaluations and student test scores in Year 2. Score gains in Year 2 are significantly and positively correlated with mentor evaluations but not with teaching fellow evaluations.
Differences based on choice of outcome	n.a.
Other key findings	With regard to subjective evaluations (mentors' or interviewers'), "there is evidence of variation in the leniency with which standards were applied by some evaluators. Specifically, variation in evaluations *within* evaluators is a much stronger predictor of student outcomes than variation *between* evaluators."

[a]The VA estimation is an intermediate step in this analysis. Teachers' VA estimates are first calculated and then used as right-hand side variables in the main equation (which evaluates the degree to which they can predict achievement gains for teachers' future students). Thus, not much detail is provided on the actual VA calculations/model itself.

VA = value added; VAM = value-added model; n.a. = not applicable.

Table A.19. Rothstein (2009)

Citation	Rothstein, J. "Student Sorting and Bias in Value-Added Estimation: Selection on Observables and Unobservables." *Education Finance and Policy,* vol. 4, no. 4, 2009, pp. 537–571.
Keywords	Bias in value-added estimates, nonrandom sorting of students
Type of VAM	Teacher (classroom)
Data	
Actual or simulated?	Actual
Location	North Carolina
Grades and subjects	5 in Reading (similar results in Math)
School years	2000–2001
Outcomes	Reading, as measured by annual end-of-year tests
Control Variables	
Student characteristics	Gender, race/ethnicity, learning disabilities in reading or in any area, Title 1 participation, each possible "exceptionality" (gifted, hearing impaired, mentally handicapped, and so on) parental years of education, free and reduced-price lunch participation, reporting never doing any homework, and number of hours of television watched each school day (only in some models)
Other controls	
School effects included?	Yes
Model	
Years of baseline test scores	One, except for VAM 4, which uses three years of baseline scores
Baseline subject controls	Reading and Math
Years of growth data	One
How does study address missing data and/or attrition	Includes only students in 5th grade in 2000–2001, with a valid teacher assignment in that year and with complete grades 3–5 test score data
Partial year enrollment or team teaching	n.a.
Adjustments (such as minimum number of students, shrinkage, or measurement error correction)	Measurement error in test scores accounted for by using a measure of "test-retest reliability"
Key Findings	If principals make classroom assignments using information about students' potential gains that are unobservable to researchers, best feasible VAMs might be substantially biased. The magnitude of the bias depends on the degree to which researchers can account for whatever assignment rule is used. If classroom assignments are random, conditional on observable variables, bias due to sorting is almost entirely removed.
Results	
Standard deviation of VAM measures	0.096–0.208 (adjusted for sampling error, if classroom assignment random conditional on observables, panel 2 of Table 5 0.100–0.114 (adjusted for sampling error, if classroom assignment is based on unobservables, Table 10)
Percentage significant effects	n.a.
Additional years of growth data	n.a.
Inter-temporal correlation	n.a.
Correlation with other measures	n.a.
Differences based on choice of outcome	n.a.
Other key findings	n.a.

VAM = value-added model.

n.a. = not applicable.

Table A.20. Rothstein (2010)

Citation	Rothstein, J. "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement." *Quarterly Journal of Economics,* vol. 125, no. 1, 2010. pp. 175–214.
Keywords	Teacher value added, teacher quality, student achievement, classroom assignment
Type of VAM	Teacher
Data	
Actual or simulated?	Actual
Location	North Carolina
Grades and subjects	Grade 5 in math and reading
School years	2000–2001
Outcomes	Math and reading end-of-grade tests
Control Variables	
Student characteristics	- Main model has three basic specifications (VAM1, VAM2, and VAM3) and one rich specification (VAM4). None include student characteristics, though VAM3 includes student indicators. - To test robustness of results, VAM1 and VAM2 are reestimated with controls for student race, gender, FRPL status, fourth-grade absences, and fourth-grade television viewing.
Other controls	VAM4 controls for teacher assignments in grades 3 and 4.
School effects included?	Yes.
Model	
Years of baseline test scores	VAM1,2, 3: One year. VAM4: Three years (from grades 2, 3, and 4). The grade 2 test is actually administered at the start of grade 3.
Baseline subject controls	VAM1, 2, and 3: Math and reading for each subject, respectively. VAM4: Both math and reading are used for each subject.
Years of growth data	One
How does study address missing data and/or attrition	Students with inconsistent longitudinal records, missing test score data, or invalid matches to a fifth-grade teacher are excluded.
Partial year enrollment or team teaching	- Sample is restricted to students in self-contained classrooms. - Sample is restricted to students who do not switch schools during the grades for which classroom assignments are controlled.
Adjustments (such as minimum number of students, shrinkage, or measurement error corr.)	- Classrooms with fewer than 12 students are excluded. - Classrooms that are the only classroom in their school for that grade are also excluded.
Key Findings	This study finds that assumptions underlying common VAMs about random classroom assignments conditional on other determinants of student achievement are substantially incorrect in North Carolina. Teacher VAM estimates based on these models cannot be interpreted as causal. Clear evidence of this is that each VAM indicates that fifth-grade teachers have quantitatively important "effects" on students' fourth-grade learning.
Results	
Standard deviation of VAM measures	0.15 (math); 0.11 (reading)
Percentage significant effects	n.a.
Additional years of growth data	As a robustness check, the author evaluates VAM analyses using data from multiple student cohorts to distinguish between permanent and transitory components of a teacher's effects and finds that the assumptions needed to avoid bias do not hold in the data.
Inter-temporal correlation	n.a.
Correlation with other measures	A teacher's first-year effect appears to be a poor proxy for his or her longer-run impact. Only one-third of teachers in the top quintile of the distribution of two-year cumulative effects are also in the top quintile of the one-year effect distribution.
Differences based on choice of outcome	VAM1–3 are rerun using original score scales or score percentiles as outcomes instead of standardized-by-grade scores. The models continue to fail falsification tests.
Other key findings	Adding controls for student characteristics has no effect on falsification test results.

FRPL = free or reduced-price lunch; VAM = value-added model.

n.a. = not applicable.

Table A.21. Tyler, Taylor, Kane, and Wooten (2010)

Citation	Tyler, John H., Eric S. Taylor, Thomas J. Kane, and Amy L. Wooten. "Using Student Performance Data to Identify Effective Classroom Practices." *American Economic Review: Papers & Proceedings,* vol. 100, May 2010, pp. 256–260
Keywords	Teacher effectiveness, classroom practices, student achievement
Type of VAM	Teacher
Data	
Actual or simulated?	Actual
Location	Cincinnati Public School
Grades and subjects	Math and reading in grades 3 through 8
School years	2000–2001 through 2008–2009
Outcomes	- Math and reading, as measured by standardized test scores from end-of-year state mandated exams
VAM Control Variables	
Student characteristics	Gender, race/ethnicity, ever retained in grade, special education, gifted, LEP
Other controls	indicator variables for the teacher's years of classroom experience; prior year scores interacted with each grade level; grade-by-year indicators
School effects included?	No
Model	
Years of baseline test scores	One
Baseline subject controls	Math and reading (for math and reading outcomes, respectively)
Years of growth data	One
How does study address missing data and/or attrition	For students with missing baseline scores, authors impute them with the grade-by-year mean and include an indicator for missing baseline scores.
Partial year enrollment or team teaching	The class schedule data retains only students' last class assignment for each course each year. This structure does not allow the authors to identify students who had more than one teacher or class during the year (or semester).
Adjustments (such as minimum number of students, shrinkage, or measurement error corr.)	Study uses a shrinkage estimator.
Key Findings	This study finds that classroom management and instructional skills, as measured by Cincinnati's TES (based on the Danielson framework), meaningfully predict student achievement growth. An overall average TES score increase of one point (that is, students assigned to a distinguished teacher versus a proficient teacher) is associated with achievement gains of 0.171 SDs in math and 0.212 SDs in reading.
Results	
Standard deviation of VAM measures	Study does not include teacher effects
Percentage significant effects	n.a.
Additional years of growth data	n.a.
Inter-temporal correlation	n.a.
Correlation with other measures	n.a.
Differences based on choice of outcome	n.a.
Other key findings	- Teachers who place a higher relative importance on the classroom environment versus exact teaching practices are predicted to raise student achievement by a greater amount (0.25 SDs in math and 0.15 SDs in reading). - Teachers who score higher on inquiry-based teaching relative to routinized standards- and content-focused teaching are predicted to produce relatively higher student gains in reading (0.150 SDs) but not in math. - Results exhibit heterogeneity across subjects and grade levels. In math, the overall TES measure predicts student achievement growth much more strongly in elementary grades 3 to 5 (0.41 SDs) than in middle school grades 6 to 8 (0.11 SDs and not significant). In reading, the coefficient on overall TES measure is somewhat larger for grades 6 to 8 (0.29 versus 0.21, both significant). (p.26 of Working Paper below).

Tyler, John H., Eric S. Taylor, Thomas J. Kane, and Amy L. Wooten. "Using Student Performance Data to Identify Effective Classroom Practices." Working Paper. December 2009.

LEP = limited English proficiency; SD = standard deviation; VAM = value-added model.

n.a. = not applicable.

THIS PAGE LEFT BLANK FOR DOUBLE-SIDED PRINTING

APPENDIX B. VAM IMPLEMENTATION IN SCHOOL DISTRICTS AND STATES

Table B.1. Dallas Independent School District

VAM's Main Function	The district uses value-added results as a component of its opt-in Professional Development and Appraisal System (PDAS) for teachers and its opt-in Principal Incentive Pay Program (PIPP). These programs provide pay-for-performance awards of up to $3,200.
Description of Composite Evaluation Measure and Component Weighting	PDAS includes the following components: (1) student achievement growth as measured by the district-created Classroom Effectiveness Index; (2) classroom observations; (3) professional development participation; and (4) teacher attendance. No information is available on weights. PIPP includes the following components: (1) State Accountability Rating [20 %]; (2) School Effectiveness Index [20 %]; (3) Texas Assessment of Knowledge and Skills (TAKS) percentage of student groups that either met the annual performance targets or had at least 90% of students that met the standard [25 %]; (4) TAKS percentage of student groups that either met the annual commended performance targets or had at least 30% of students (50% for magnets) with TAKS scale scores of at least 2,400 points [10 %]; (5) On-Track Performance: percentage of student groups that either met the annual performance targets or had at least a 93% graduation rate or had at least 90% of students that met the standard [25%].
Included Grades, Subjects, and Assessments	Norm-referenced math and reading tests (grades 1–9); TAKS in reading and math (grades 3–11), writing (grades 4 and 7), science (grades 5,10, and 11), and social studies (grades 8, 10, and 11); assessments of course performance in language arts, math, social studies, science, foreign language, and technology (grades 7–12).
Staff in Nonincluded Grades or Subjects	Teachers in subjects without standardized or valid course measures are initially excluded from the program. They will be included as reliable and valid measures of the performance of student performance are added to the value-added system.
School-Based Incentive Involving Value-Added	The pay-for-performance system includes compensation for school-wide improvement. The Outstanding School Performance Award rewards principals and full-time professional or support personnel (part-time on prorated basis). The basis for the award is student achievement status and VA, measured primarily by the TAKS but also by other tests.

Sources:

Dallas Independent School District. "Dallas ISD Principal and Teacher Incentive Fund Proposal." Retrieved from http://www.dallasisd.org/performancepay/incentive/publications/Teacher_Incentive_Pay_Model_final_brief.doc on August 24, 2010.

Dallas Independent School District. "MyData Portal." Retrieved from https://mydata.dallasisd.org/ on August 24, 2010.

Dallas Independent School District. "Outstanding School Performance Awards 2005-06." Retrieved from http://www.dallasisd.org/performancepay/incentive/publications/SchoolPerformanceAwardsManual_200506.pdf on August 24, 2010.

National Comprehensive Center for Teacher Quality. "Guide to Teacher Evaluation Products." Retrieved from http://www3.learningpt.org/tqsource/GEP/GEPTool.aspx?gid=87&tid=8 on August 24, 2010.

VA = value added; VAM = value-added model.

Table B.2. Florida

VAM's Main Function	Florida requires that teacher evaluations rely on classroom observations and objective measures of student learning, including state assessment data. However, the state does not use its data system to provide value-added evidence of teacher effectiveness. In April 2010, Governor Crist vetoed a bill that would have eliminated tenure for new teachers and required that value added be used in teacher evaluations.
Description of Composite Evaluation Measure and Component Weighting	The proposed evaluation system that Governor Crist vetoed would have given a 50 percent weight to value-added scores. The current evaluation instruments include classroom observations and objective evidence of improved student learning, as measured by state assessment data or, for grades or subjects not tested by the state assessment, local assessments and peer evaluations. Other criteria include a teacher's knowledge and skills and a demonstrated ability to maintain discipline.
Included Grades, Subjects, and Assessments	The state administers the Florida Comprehensive Assessment Test (FCAT) to students in grades 3 to 11 in reading, writing, math, and science.
Staff in Nonincluded Grades or Subjects	n.a.
School-Based Incentive Involving Value-Added	Florida's A+ Plan is a school-wide incentive program but it does not involve a value-added model. It grades each school in the state on an A-to-F basis, based on student performance and progress on the FCAT. Schools receiving an A or that improve a letter grade are eligible for additional funding of $100 per student. Schools can use these funds for a variety of purposes, including teacher bonuses and school improvements. Schools rated as grade F receive substantially more money per student in improvement assistance than do A-rated schools.

Sources:

Christ, Charlie. "A+ Plan for Education." Retrieved from http://www.flgov.org/a_plus_plan on August 24, 2010.

Florida Department of Education. "District Performance Appraisal System Checklist." Retrieved from http://www.fldoe.org/profdev/pa.asp on August 24, 2010.

Miami Herald. "Gov. Charlie Crist vetoes Florida teacher pay bill; what happens next?" April 16, 2010. Retrieved from http://www.miamiherald.com/2010/04/16/1582150/why-charlie-crist-vetoed-the-teacher.html on August 24, 2010.

National Council on Teacher Quality. "2009 State Teacher Policy Yearbook: Florida report." Retrieved from http://www.nctq.org/stpy09/ on August 24, 2010.

n.a. = not applicable.

Table B.3. Louisiana

VAM's Main Function	In May 2010, Governor Jindal signed into law a bill requiring that teacher and school administrator evaluations include evidence of student achievement growth, as measured by value-added assessment. VAM-based estimates will also be required for evaluations of school administrators. The new evaluation system will be piloted in two dozen volunteer school districts during the 2010–2011 and 2011–2012 school years. Statewide implementation will take place in 2012–2013. Teaching evaluations will be used for targeting professional development and for recertification purposes. Teachers rated as ineffective will receive additional professional development. They will not be recertified with three ineffective ratings during their certification cycle unless the school board appeals. School districts will maintain control over how teachers are compensated, rewarded, and retained.
Description of Composite Evaluation Measure and Component Weighting	Teacher evaluations will give a 50 percent weight to student achievement growth, as measured by value added, and a 50 percent combined weight to principal observations, peer reviews, and other subjective criteria.
Included Grades, Subjects, and Assessments	The value-added model will include the Louisiana Education Assessment Program (LEAP) and other (unspecified) testing data.
Staff in Nonincluded Grades or Subjects	The State Board of Elementary and Secondary Education will adopt policies to measure student growth in grades and subjects in which value-added data are not available. These policies will be informed by recommendations from an advisory committee that includes practicing educators.
School-Based Incentive Involving Value-Added	Beginning in 2011–2012, school performance scores will also include a value-added component. The existing measure includes information on test score levels, attendance, graduation rates, and dropout rates.

Sources:

Louisiana Advocate Capitol News Bureau. "Jindal signs teacher evaluation measure." May 28, 2010. Retrieved from http://www.2theadvocate.com/news/95085169.html?showAll=y&c=y on August 24, 2010.

Louisiana Department of Education. "Louisiana adopts value-added teacher evaluation model." Retrieved from http://doe.louisiana.gov/lde/comm/pressrelease.aspx?PR=1428 on August 24, 2010.

Office of the Governor of Louisiana. "Governor Jindal signs groundbreaking teacher evaluation bill into law." Retrieved from http://gov.louisiana.gov/index.cfm?md=newsroom&tmp=detail&catID=2&articleID=2200 on August 24, 2010.

VAM = value added model.

Table B.4. Memphis City Schools

VAM's Main Function	Memphis' Teacher Effectiveness Measure (TEM) enables Memphis City Schools to make decisions about tenure, dismissal, compensation, retention bonuses, and differentiated roles. This is part of a Teacher Effectiveness Initiative that is being piloted during the 2010–2011 school year.
Description of Composite Evaluation Measure and Component Weighting	The TEM consists of the following measures: growth in student learning (VAM-based) [35%], observation of teachers' practice [35%], perceptions by stakeholders (students, parents, colleagues) [15%], and teacher content and pedagogical knowledge [15%].
Included Grades, Subjects, and Assessments	Value-added data will come from TVAAS.
Staff in Non-Included Grades or Subjects	Teachers without value-added scores will have the weights of the other three TEM components increased proportionately. Memphis plans additional assessments to expand value added to the majority of teachers in all core content areas. The plan envisions having value-added measures for 65 percent of MCS teachers, an increase of 35 percentage points. Memphis will also convene a task force to develop an approach to capturing value-added data for specialist teachers (for example, special education, English as a second language, or reading specialists).
School-based Incentive Involving Value-Added	Teachers can earn a performance-based group bonus for achieving student growth goals. Memphis will pilot group bonus opportunities for various types of "teams" (for example, the kindergarten through grade 3 continuum, upper elementary grade levels, secondary-level content areas). Group bonus awards are expected to be in the range of $2,500 per teacher.

Source: Memphis City Schools. Teacher effectiveness homepage. Retrieved from http://www.mcsk12.net/tei/index.asp on August 24, 2010.

TVAAS = Tennessee value-added assessment system; VAM = value-added measure.

Table B.5. North Carolina

VAM's Main Function	North Carolina's teacher and principal evaluation standards, adopted in 2007, do not include a value-added component. All schools implemented the current evaluation standards for the first time in 2010–2011 after piloting in 2009–2010. The North Carolina Education Value-Added Assessment System (EVAAS) provides schools and districts with grade-level reports on the trajectory of cohort and student subgroup achievement gains and with diagnostic reports to identify students at risk for underachievement. Educators use EVAAS primarily to tailor instruction in ways that meet each student's academic needs. However, EVAAS, or any measure of student growth, is not currently part of teacher or principal evaluations.
Description of Composite Evaluation Measure and Component Weighting	Teacher evaluation components include self-assessments, teacher-principal conferences, classroom observations, and professional development plans. Teachers are evaluated on five standards: demonstrates leadership, establishes a respectful environment for diverse students, knows the content, facilitates learning for students, and reflects on practice. Principal evaluation components include self-assessments and input from various stakeholders. Principals are evaluated according to seven standards: strategic leadership, instructional leadership, cultural leadership, human resource leadership, managerial leadership, external development leadership, and micro-political leadership.
Included Grades, Subjects, and Assessments	The NC EVAAS includes math and reading comprehension (grades 3–8), and science (grades 5 and 8) as measured by end-of-grade assessments. End-of-course assessments are used for algebra I, algebra II, English I, U.S. history, civic and economics, biology, and physical science.
Staff in Nonincluded Grades or Subjects	n.a.
School-Based Incentive Involving Value-Added	n.a.

Sources:

National Council on Teacher Quality. "2009 state teacher policy yearbook: North Carolina report." Retrieved from http://www.nctq.org/stpy09 on August 24, 2010.

Public Schools of North Carolina. "North Carolina principal evaluation process." Retrieved from http://www.dpi.state.nc.us/docs/profdev/training/principal/principal-evaluation.pdf on August 24, 2010.

Public Schools of North Carolina. "North Carolina teacher evaluation process." Retrieved from http://www.dpi.state.nc.us/docs/profdev/training/teacher/teacher-eval.pdf on August 24, 2010.

VAM = value added model.

n.a. = not applicable.

Table B.6. Pennsylvania

VAM's Main Function	The Pennsylvania Value Added Assessment System (PVAAS) provides schools and school districts with grade-level reports on the trajectory of cohort achievement gains. Educators use PVAAS for professional development purposes, but PVAAS, or any measure of student growth, is not currently part of teacher or principal evaluations. The commonwealth is currently piloting the development of a new model evaluation system for teachers and principals that will include student growth as a significant factor. Similar efforts are underway in Pittsburgh Public Schools as well.
Description of Composite Evaluation Measure and Component Weighting	n.a.
Included Grades, Subjects, and Assessments	PVAAS includes math and reading (grades 3–8 and 11), science (grades 4, 8, and 11), and writing (grades 5, 8, and 11) as measured by the Pennsylvania System of School Assessments.
Staff in Nonincluded Grades or Subjects	n.a.
School-Based Incentive Involving Value-Added	n.a.

Sources:

Governor of Pennsylvania Website. "Pennsylvania Value Added Assessment System (PVAAS)." Retrieved from http://www.governor.state.pa.us/portal/server.pt/community/pa_value-added_assessment_system_%28pvaas%29/8751 on August 24, 2010.

Pennsylvania Website. "PVAAS District Considerations for Implementation & Planning." Retrieved from http://www.portal.state.pa.us/portal/server.pt/gateway/PTARGS_0_123031_890978_0_0_18?PVAAS_Implementation_and_Planning_for_Districts.pdf on August 24, 2010.

VAM = value added model.

n.a. = not applicable.

Table B.7. Tennessee

VAM's Main Function	The Tennessee First to the Top Act of 2010 introduced a new evaluation system for teachers and principals that will use student growth as measured by the Tennessee Value Added Assessment System (TVAAS) as a significant factor. The state has long been collecting longitudinal information on students and teachers and using TVAAS for school and teacher value-added reporting. The new system strengthens the role of value added in teacher and principal evaluations. Beginning in fiscal year 2011, the act requires annual evaluation of all teachers and principals and that personnel decisions (including promotion, retention, tenure, and compensation) be based in part on these evaluations. A newly created Teacher Evaluation Advisory Committee is charged with developing and recommending to the State Board of Education guidelines and criteria for the annual evaluation process. Recommendations for the new evaluation framework are scheduled for November 2010.
Description of Composite Evaluation Measure and Component Weighting	TVAAS is one component of the composite evaluation measure for teachers and principals. For teachers with value-added scores, the components and weights include TVAAS [35%]; student performance on assessments, end-of-year subject tests, AP exams, and so on [15%]; and subjective measures [50%], such as conferences about strengths and weaknesses, classroom observations, and written performance assessments. For principals, the criteria can include additional factors pursuant to their employment contracts. School value added from TVAAS is used instead of teacher value-added scores for principal evaluations.
Included Grades, Subjects, and Assessments	TVAAS includes the performance of students in grades 3 to 8 on the Tennessee Comprehensive Assessment Program, a series of assessments in reading, language arts, math, science, and social studies. At the high school level, it includes state end-of-course tests in algebra I, biology, and English II.
Staff in Nonincluded Grades or Subjects	For teachers with no TVAAS data, other comparable measures of student growth can be used (35% weight). There are currently two options for teachers in untested grades: (1) school value added is the entire 35%, the default; (2) school value-added is 20% with the remaining 15% developed by the district. Option two applies to teachers in untested subjects. School value added will factor into evaluations for librarians and other staff, but the other components are not yet finalized.
School-Based Incentive Involving Value-Added	TVAAS provides school reports on value added.

Source: Tennessee First to the Top website. Retrieved from http://www.tn.gov/firsttothetop/programs.html on August 24, 2010.

VAM = value-added measure.

THIS PAGE LEFT BLANK FOR DOUBLE-SIDED PRINTING

APPENDIX C. TECHNICAL DESCRIPTION OF VALUE-ADDED MODELS

A. Teacher Value-Added Models

The basic VAM for estimating teacher effects predicts the achievement scores of individual students using factors such as prior test scores, student and peer background characteristics that are thought to correlate with achievement, and a student's teacher. The estimation equation is

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

where $A_{i,j,y}$ is the achievement score for student i in subject j (for example, math or reading) in year y and $A_{i,j,y-1}$ is the prior score for student i in subject j. Analysts control for prior scores differently. Mathematica typically includes prior test scores in all available subjects. $X_{i,j,y}$ is a set of control variables for student and classroom characteristics. The components of X vary across estimation models, but analysts commonly include factors such as gender, meal program status, race/ethnicity, disability, gifted program participation, and grade level. $D_{i,j,y}$ is a set of teacher variables, and $e_{i,j,y}$ is the error term. The coefficients β, γ, and δ capture the estimated relationships between the outcome variable and each respective component in the model.

Mathematica uses a dosage approach in constructing the teacher variables as a flexible way of accounting for students who learn from more than one teacher in a given subject in a year. The vector $D_{i,j,y}$ includes one variable for each teacher. Each variable equals the fraction of the year student i was taught by a teacher in subject j. The dosage value of any element of $D_{i,j,y}$ is zero if student i was not taught by that teacher. We also include a residual teacher dosage term that equals $1-\Sigma_T D_T$, where $\Sigma_T D_T$ for a given student is the sum of dosage variables across all teachers. For some students, we have information on their teachers for only part of the year. The residual teacher dosage variable accounts for the remaining time. These measures are elements of the vector δ, which are the coefficients on $D_{i,j,y}$.

In constructing the model, Mathematica normalizes each term by subtracting its mean and dividing by its standard deviation within each grade and year. This process allows information from multiple grades to be meaningfully included together. After estimating the VAM, we use a shrinkage procedure to minimize the possibility that teachers with relatively few students are overrepresented among high- and low-performing teachers. Finally, we center the teacher estimates on a zero value. A teacher with a VAM estimate of zero is contributing the average amount to student achievement growth among teachers in the sample.

Analysts can then examine several alternative model specifications, such as models that combine multiple student cohorts (that is, multiple years of teaching), or models that include control variables for multiple years of students' own test scores (that is, students' test scores from the prior-prior years in addition to their scores from the prior year). Analysts can also explore including school dosage variables with the teacher dosage variables in the VAM. Including school effects provides a more flexible set of background controls but it also fundamentally changes the inference to a comparison of teacher effectiveness within each school rather than across the entire sample.

B. School Value-Added Models

The VAM for estimating school effects closely resembles the teacher model. In fact, it is operationally the same except that the school VAM includes school dosage variables instead of

teacher dosage variables. The school dosage variables are calculated similarly to the teacher dosage variables. Each variable equals the fraction of the year student *i* was taught at a given school. The dosage value of any element of $D_{i,y}$ is zero if student *i* was not taught at that school. Like the teacher VAM, we also include a residual school dosage term that equals $1-\Sigma_S D_S$, where $\Sigma_S D_S$ for a given student is the sum of dosage variables across all schools. The residual school dosage variable accounts for the remaining time.

Like the teacher VAM, the school model incorporates only factors that are observable in the data and assumes that students' prior test scores control for their achievement trajectories. A limitation of the school model is that (unlike in the teacher model) a student's score in the previous year is typically not a pretreatment baseline score. Students are generally served by the same school both in the current year (that is, the year to which a set of VAM estimates apply) and in the prior year, when baseline scores are measured. Consequently, students' baseline scores are not predetermined for schools as they are for teachers. The school model has a related limitation in that estimates can only include tested grades but they apply to all grades at the school. This limitation means that, if assessments start in third grade, as statewide assessment does in Pennsylvania, school VAM scores for elementary schools will include only grades 4 and 5. For middle schools, however, they can include the entire grade 6 to 8 span.

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

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