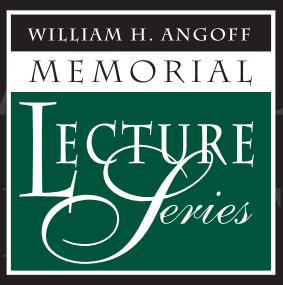
ETS®

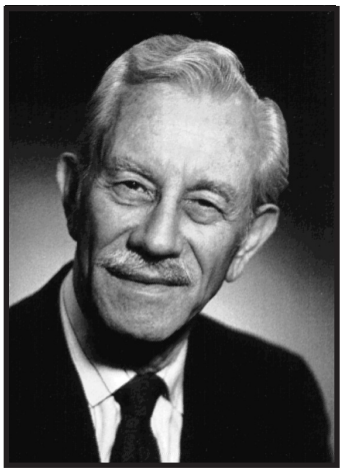
Listening. Learning. Leading.®

RELIABILITY AND VALIDITY OF INFERENCES ABOUT TEACHERS BASED ON STUDENT TEST SCORES

By Edward H. Haertel



William H. Angoff
1919–1993

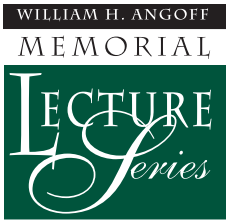


William H. Angoff was a distinguished research scientist at ETS for more than 40 years. During that time, he made many major contributions to educational measurement and authored some of the classic publications on psychometrics, including the definitive text "Scales, Norms, and Equivalent Scores," which appeared in Robert L. Thorndike's Educational Measurement. *Dr. Angoff was noted not only for his commitment to the highest technical standards but also for his rare ability to make complex issues widely accessible.*

The Memorial Lecture Series established in his name in 1994 honors Dr. Angoff's legacy by encouraging and supporting the discussion of public interest issues related to educational measurement. These lectures are jointly sponsored by ETS and an endowment fund that was established in Dr. Angoff's memory.

The William H. Angoff Lecture Series reports are published by the Center for Research on Human Capital and Education, ETS Research and Development.

Reliability and Validity of Inferences About Teachers Based on Student Test Scores



The 14th William H. Angoff Memorial Lecture was presented at The National Press Club, Washington, D.C., on March 22, 2013.

Edward H. Haertel
Stanford University

ETS
Research & Development
Center for Research on Human Capital and Education
Princeton, NJ 08541-0001

Preface

The 14th William H. Angoff Memorial Lecture was presented by Dr. Edward H. Haertel, Jacks Family Professor of Education, Emeritus, Stanford University. In his lecture, Dr. Haertel examines the use of value-added models (VAM) in measuring teacher effectiveness. VAMs, complex statistical models for calculating teacher value-added estimates from patterns of student test scores over time, have been receiving increasing attention as a method for states to revise or establish teacher evaluation systems to take into account the effect of individual teachers on student achievement. These models provide scores for teachers, intended to tell how well each did in raising achievement of their students. Using a test validation methodology in assessing VAMs, Haertel examines questions of validity, reliability, prediction power, and potential positive and negative effects of particular uses of teacher value-added scores. His lecture, which includes cautionary notes about using value-added scores in making high-stakes decisions, adds to the public policy discussion of teacher performance evaluation methods.

The William H. Angoff Memorial Lecture Series was established in 1994 to honor the life and work of Bill Angoff, who died in January 1993. For more than 50 years, Dr. Angoff made major contributions to educational and psychological measurement and was deservedly recognized by the major societies in the field. In line with Dr. Angoff's interests, this lecture series is devoted to relatively nontechnical discussions of important public interest issues related to educational measurement.

Ida Lawrence
Senior Vice President
ETS Research & Development
September 2013

Acknowledgments

My thanks go to Robert Mislevy and to Ida Lawrence for the invitation to deliver the 14th William H. Angoff Memorial Lecture, presented March 21, 2013, at ETS in Princeton, New Jersey, and the following day at the National Press Club in Washington, D.C. It has been revised slightly for publication. I am most grateful for thoughtful and constructive comments from several colleagues and reviewers along the way, including Derek Briggs and Jesse Rothstein for their review of an early draft, as well as James Carlson, Daniel McCaffrey, Gary Sykes, and others for their helpful comments on a later version. Their help has been invaluable both in preparing the original talk and in revising it for publication. The views expressed are mine alone, of course, and I am entirely responsible for any remaining errors. Kimberly Ayotte provided outstanding logistical support of all kinds, especially when the lectures had to be cancelled due to Hurricane Sandy and then rescheduled. James Carlson, Richard Coley, and Kim Fryer have provided superb editorial assistance.

Abstract

Policymakers and school administrators have embraced value-added models of teacher effectiveness as tools for educational improvement. Teacher value-added estimates may be viewed as complicated scores of a certain kind. This suggests using a test validation model to examine their reliability and validity. Validation begins with an interpretive argument for inferences or actions based on value-added scores. That argument addresses (a) the meaning of the scores themselves — whether they measure the intended construct; (b) their generalizability — whether the results are stable from year to year or using different student tests, for example; and (c) the relation of value-added scores to broader notions of teacher effectiveness — whether teachers' effectiveness in raising test scores can serve as a proxy for other aspects of teaching quality. Next, the interpretive argument directs attention to rationales for the expected benefits of particular value-added score uses or interpretations, as well as plausible unintended consequences. This kind of systematic analysis raises serious questions about some popular policy prescriptions based on teacher value-added scores.

Introduction

It seems indisputable that U.S. education is in need of reform. Elected officials, school administrators, and federal policymakers are all frustrated with achievement gaps, vast numbers of schools in need of improvement under the No Child Left Behind Act (NCLB, 2002), and a drumbeat of bad news comparing U.S. test scores to those of other nations. It seems we hear daily about declining college and career readiness, 21st-century skills, and global competitiveness if public education does not improve.

At the same time, the belief has spread that research shows just having a top quintile teacher versus a bottom quintile teacher for 5 years in a row could erase the Black-White achievement gap (Ravitch, 2010).

It is also widely recognized that our ways of identifying and dismissing poor-performing teachers are inadequate, that teacher credentials alone are poor guides to teaching quality, and that teacher evaluation in most school districts around the country is abysmal.

What could be more reasonable, then, than looking at students' test scores to determine whether or not their teachers are doing a good job? The teacher's job is to teach. Student test scores measure learning. If teachers are teaching, students should learn and scores should go up. If they are teaching well, scores should go up a lot. If test scores are not moving, then the teachers should be held accountable.

There are some messy details, of course, in translating student test scores into teacher effectiveness estimates, but sophisticated statistical models, referred to as *value-added models* (VAMs), have been created to do just that. Dozens of highly technical articles in leading journals are devoted to these models; data systems linking student test scores over time to individual teachers have improved enormously in recent years. It seems the time has come. Common sense and scientific research both seem to point to teacher evaluation based on VAMs as a powerful strategy for educational improvement.

In this lecture, I first comment on the importance of teacher effectiveness and the argument concerning top quintile teachers. I next turn to the importance of sound test score scales for value-added modeling, followed by the logic of VAMs and the statistical challenges they must overcome. The major portion of these remarks is devoted to describing an *interpretive argument* (Kane, 2006) for teacher VAM scores and the associated evidence. The interpretive argument is essentially a chain of reasoning from the construction of teacher VAM scores to the inferences those scores are intended to support. This framework is useful in organizing the many different assumptions required to support inferences about comparisons of individual teachers' effectiveness based on their students' test scores. Finally, I comment briefly on what I believe are more appropriate uses of teacher VAMs and better methods of teacher evaluation.

The Angoff Lectures are intended to be relatively nontechnical discussions. I have tried to explain VAMs in terms that any reader with a little patience should be able to follow, but I am afraid a few technical terms will be unavoidable.

Most of this lecture is concerned with the suitability of VAMs for teacher evaluation. I believe this use of VAMs has been seriously oversold, and some specific applications have been very unwise.[1] I should state at the outset, however, that like most statistical tools, these models are good for some purposes and not for others. In my conclusions, I comment briefly on what I regard as sound versus unsound uses.

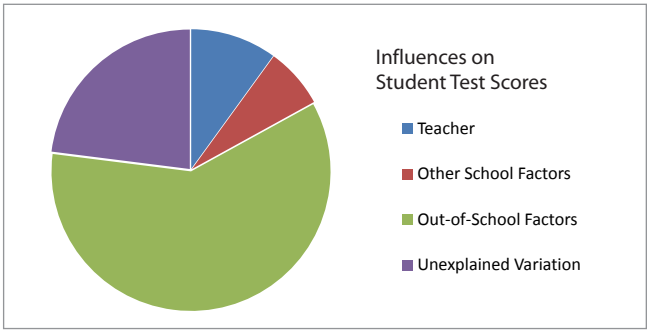
[1] See, for example, Winerip (2011).

How Much Does Teacher Effectiveness Matter?

Before getting into the details of VAMs and how they work, let us consider just how much differences in teacher effectiveness really matter for schooling outcomes. Obviously, *teachers* matter enormously. A classroom full of students with no teacher would probably not learn much — at least not much of the prescribed curriculum. But the relevant question here is how much does *variation* among teachers matter for schooling outcomes? The relevant comparison is not between some teacher and no teacher, but rather between a good teacher in some sense and a poor teacher. Teachers appear to be the most critical *within-school* influence on student learning, but *out-of-school* factors have been shown to matter even more. One recent study put the influence of out-of-school factors at 60% of the variance in student test scores, and the influence of teachers at around 9% (Goldhaber, Brewer, & Anderson, 1999).[2] Another study, using the Tennessee STAR data, found that teachers accounted for about 13% of the variance in student mathematics test score gains and about 7% of the variance in reading test score gains (Nye, Konstantopoulos, & Hedges, 2004). Some variation is always left unexplained by these models — we might refer to it as random variation or random error, but all that really means is that it is not attributable to any of the factors included in a particular model. So let us just say teacher differences account for about 10% of the variance in student test score gains in a single year.

As shown in Figure 1, whether 10% is a little or a lot depends on how you look at it. Policymakers who seek to improve schooling outcomes have to focus on potentially *changeable* determinants of those outcomes. Family background, neighborhood environment, peer influences, and differences in students' aptitudes for schooling are seen as largely beyond the reach of educational policy. Relative to just the smaller set of variables that education policies might directly influence, differences in teacher effectiveness appear quite important. In this respect, 10% may seem large. Some proportion of that 10% will remain outside the reach of policy, but on the other hand, cumulative achievement boosts year after year could add up to a somewhat larger effect. However, if the goal is to dramatically change patterns of U.S. student achievement, then identifying and removing low-performing teachers will not be nearly enough. As my colleague Linda Darling-Hammond has quipped, "You can't fire your way to Finland" ("An Education Exchange," 2011, Teaching Quality Partnerships section, para. 8).

Figure 1
How Much Variance in Student Test Score Gains Is Due to Variation Among Teachers?



There is another sense in which 10% is small. It is small relative to the 90% of the variation due to other factors, only some of which can be explained. Simply put, the statistical models used to estimate teacher VAM scores must separate a weak signal from much noise and possible distortion. Models can filter out much of the noise, but in the end, there is still much remaining.

[2] Goldhaber et al. (1999) reported that roughly 60% of variance in test scores is explained by individual and family background variables, which included a prior year test score.

The Myth of the Top Quintile Teachers

I mentioned the often-repeated story that a string of top quintile teachers versus bottom quintile teachers could erase the Black-White achievement gap in 5 years. Some researchers have suggested 4 years, others 3 years (Ravitch, 2010, pp. 181 ff.). Where do these numbers come from? If test score gains are calculated for every student — just this year's score minus last year's score — and then averaged up to the teacher level, an average test score gain can be obtained for each teacher. (Actual procedures are more complicated, but this will work as a first approximation.) Next, the one fifth of the teachers with the highest average gains can be compared to the one fifth with the lowest gains. The gap between the means for those two groups may be termed the *effect* of having a top quintile teacher versus a bottom quintile teacher. Suppose that comes out to 5 percentile points. If the Black-White achievement gap is 25 percentile points, then one could claim that if a student got a 5-point boost each year for 5 years in a row, that would be the size of the gap. This sounds good, but there are at least three reasons why such claims may be exaggerated.

Measurement Error

Number one, it is not certain who those top quintile teachers really are. Teacher value-added scores are unreliable. As will be shown, that means the teachers whose students show the biggest gains one year are often not the same as those whose students show big gains the next year. Statistical models can do much better than chance at predicting which teachers' students will show above-average gains, but these predictions will still be wrong much of the time. If one cannot be confident about *which* teachers are the top performers, then the full benefit implied by the logic of the top quintile/bottom quintile argument cannot be realized.

Measurement error will lead to unrealistically large teacher-effect estimates if the very same student test scores used to calculate teacher value-added are then used again to estimate the size of the teacher effect. This incorrect procedure amounts to a circular argument, whereby highly effective teachers are defined as those producing high student test score gains and those same students' test score gains are then attributed to their having been assigned to highly effective teachers. If a study instead classifies teachers into quintile groups based on their students' performance one year and then examines the performance of *different* students assigned to those teachers in a later year, the estimated quintile effect should correctly incorporate the effects of measurement error.[3]

Perhaps the first top quintile claim to attract widespread media attention was a study by Sanders and Rivers (1996). Using data from two urban school districts in Tennessee, these authors predicted a 50 percentile point difference between students assigned to top quintile versus bottom quintile teachers for 3 years in a row. Although the description of their statistical model is incomplete, it appears that measurement error may have led to an inflated estimate in this study, and that their finding was probably overstated (Kupermintz, Shepard, & Linn, 2001).

[3] Because teachers' estimated value-added scores always include some measurement error, teachers classified as top quintile or bottom quintile are not truly the most or the least effective. Random error causes some mixing of less effective teachers into the top group and more effective teachers into the bottom group. Thus, teachers *classified* as top quintile or bottom quintile do not all truly belong in those respective groups, and the effect estimated on the basis of teacher *classifications* will be smaller than the hypothetical effect attributable to their (unknown) true status.

Fade-Out

Problem number two has to do with the idea that one can simply add up gains across years to get a total effect. In fact, the effects of one year's teacher, for good or for ill, fade out in subsequent years. The effects of that wonderful third grade teacher will be much attenuated by the time a student reaches seventh grade. So, the cumulative effect of a string of exceptional teachers will be more than the single year effect, but considerably less than a simple summation would imply.

Teacher effects do not fade out entirely, of course. In a recent study, Chetty, Friedman, and Rockoff (2011) estimated that about 30% of the teacher effect persists after 3 or 4 years, with little further decline thereafter. They report that this is generally consistent with earlier research, but they are able to provide more accurate and longer term estimates using an exceptionally large longitudinal data set. Their study is also exceptional in detecting elementary school teacher effects lasting even into young adulthood.

Implementation Challenge

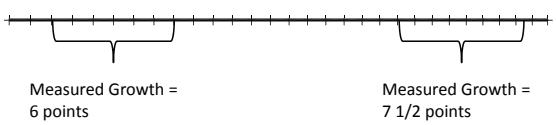
Finally, problem number three is simply that there is no way to assign all of the top performing teachers to work with minority students or to replace the current teaching force with all top performers. The thought experiment cannot be translated into an actual policy.

Calculating Test Score Gains

I glossed over another challenge in the top quintile/bottom quintile story when I began with test score gains calculated for each student by simply subtracting last year's score from this year's score. Measuring student achievement is not the same as measuring length or weight. The difference between 2 inches and 4 inches is the same as the difference between 6 inches and 8 inches. That is what is meant by an *equal-interval scale*. But, it is much harder to be sure that the difference between test scores of 20 and 40 is the same as the difference between scores of 60 and 80. Notice I did *not* refer to getting 20 items right or 40 items right. Raw scores are pretty much hopeless for these purposes. Test developers use sophisticated statistical models to convert raw scores into *scale scores* with better statistical properties, but these scale scores are still far from perfect. What does equal interval mean in describing test score scales? Does it mean that on average, it takes the same amount of instructional time or teaching skill to boost a student's score from 20 to 40 as it does from 60 to 80? Probably not, actually. The short answer is that the meaning of equal interval varies according to the score scale's intended use or interpretation, and even for a specific intended use, whether or not a scale is equal interval cannot generally be determined.

So why does having an equal interval scale matter? Let us say the score scale is not equal interval. To take just one possible example, let us suppose the units near the top of the scale are actually a little bit smaller than at the bottom of the scale. In that case, as shown in Figure 2, if two teachers' students start out at different score levels, on average, and if the teachers would in fact appear equally effective in raising student test scores on an equal-interval scale, then the *measured* gains for the students in the higher-performing classroom will appear larger. A direct comparison of measured score gains for the two teachers will be unfair.[4]

Figure 2
Possible Consequences of a Nonlinear Test Score Scale



A nonlinear scale means teachers are rewarded or penalized, depending on where their students start out

Especially problematical for teachers of students above or below grade level, or with special needs

This is not just a hypothetical argument. Tests aligned to grade-level standards cannot fully register the academic progress of students far above grade level or far below grade level. If the test is too hard for the students, then they may make much progress and still score around the chance level. And if the test is too easy, students may get near-perfect scores on the pretest and not do much better when they are tested again a year later. That translates into bias against those teachers working with the lowest-performing or the highest performing classes.[5] If tests have an inadequate range of content and difficulty, then bias against some teachers is likely.

[4] The statistical challenge is even greater when the equal-interval scale has to span multiple grade levels. If students' gains are calculated by subtracting prior year test scores from current year scores, then these gains are probably comparisons between scores on two different tests, built to measure different grade-level content standards. Calculating gain scores across years requires something called a *vertical scale*. If tests are not vertically scaled across grades, VAMs cannot rely on gain scores and must instead incorporate prior year test scores in much the same way as any other predictor. This is satisfactory, but to the extent that prior year scores and current-year scores measure different constructs, accuracy will suffer. Note that the equal interval scale assumption is important *whether or not* a vertical scale is assumed.

[5] Teachers of high-performing students typically earn above average value-added scores, but anecdotal reports suggest that teachers of gifted and talented classrooms may be penalized because tests are too easy to measure their students' progress (Amrein-Beardsley & Collins, 2012).

The Logic of Value-Added Models

In examining the logic of VAMs, it is helpful to begin by considering briefly what is wrong with evaluating teachers just by comparing their students' average test scores at the end of the year. Seeing the obvious flaw in that approach should help to clarify the problem the VAM has to solve.

Let us begin with something familiar. Think about a typical testing situation, where each student gets a test consisting of a collection of items. The student answers the items; the answers are scored; the item scores are summed to get a test score; and finally, different students' test scores are compared to see who ranks high and who ranks low.

Next, apply this template to the problem of measuring teacher effectiveness. The comparison is shown in Table 1. This time, think about a testing situation where the examinees are *teachers*, not students, and where each test item, if you will, is actually a *student*. The way these student-items are administered to the teacher-examinees is by having the teacher teach the student for a year. The way the student-items are scored is by giving each student an achievement test at the end of the year. The way the student-item scores are summarized is by averaging the students' test scores within each classroom. Then, the teachers are compared to one another based on these averages.

Now, one can see right away that this is not going to work very well because some teachers will get students who are easier to teach or who know more at the start of the year compared to other teachers. If the group of students in a teacher's classroom for a year is like a test for that teacher, then one might say that some teachers are handed much easier tests, and others are handed much harder tests.

So to make the teacher comparisons fairer, one has to adjust for these student differences. This is done by estimating what score each student would have earned, on average, if that student had been taught all year by *any other* teacher. Then, by comparing the student's *actual* end-of-year score to this *estimated* score average across *all possible* teachers, one can adjust for those differences in the students assigned to different teachers. The starting premise is that each student spent the year being taught by one particular teacher. The end-of-year scores that would have been observed if that student had instead been taught by some other teacher are each referred to as

Table 1
Test Scores for Students Versus Value-Added Model (VAM) Scores for Teachers

Aspect of testing situation	Typical test	Simplified teacher VAM
Examinees	Students	Teachers
Items	Test questions	Students
Test	Items in a test form	Students in a classroom
Administration	Student answers items	Teacher teaches students
Item scoring	Item responses scored according to key	Student learning scored by giving each student a standardized test
Test score	Sum of item scores	Average of student test scores

counterfactuals — the hypothesized outcomes of events that did not actually happen. The student's average score across all these potential, counterfactual teacher assignments is used as the point of comparison for judging the *actual* score obtained after the student has spent the year with a *particular* teacher.

Once these counterfactual scores are estimated for each student, one can see whether each student actually performed as well as, better than, or worse than predicted. The estimated average score is subtracted from the observed score, so that a positive difference means better than expected; a negative difference means worse than expected. Then these differences are averaged up to the teacher level.[6]

So how does one estimate how well a given student would have scored after spending the year in some other teacher's classroom? One looks for students *similar* to that given student and assumes that the average observed score for those other students, obtained after their respective years of instruction with various other teachers, gives a good estimate of the average counterfactual score for the given student. Various kinds of information about students can be used in deciding what *similar* means here.

Now this is not the way VAMs are typically described. In practice, to carry out this process of estimating average counterfactual scores for each student, one makes strong statistical assumptions about the functional form of the relationships among various observable student characteristics and achievement test scores — whether relationships between variables are best characterized as linear, for example, or with some more complicated mathematical function. Then a technique called *regression analysis* is used to carry out the estimation for all students at once. The process is often described in the convenient shorthand of controlling for or adjusting for various factors. That language is perfectly fine, of course, but may make it too easy to ignore the underlying logic of the estimation and the strong assumptions the regression model actually entails.

Some big differences among various VAMs stem from their choices as to what information to use in controlling or adjusting for student differences. Prior year test scores are included, because these are among the most powerful predictors of current-year test scores. Students who scored high last year are likely, on average, to score high again this year. Of course, just looking at last year's test score is not enough. VAMs that reach back further in time, including test scores from 2 years earlier as well as from the previous year, are considerably more accurate. Some models just use prior scores from the same subject area, while others pull in test scores from different subject areas. In addition to test scores, some models use students' absences, suspensions, grade retentions, English learner or special education status, or summer school attendance. Some models may include gender or other demographic variables describing students. Models may include the average scores of other students in the same classroom or the average score for the entire school or district. All of these choices influence the resulting estimates of how well each individual student would have fared, averaging across all possible teacher assignments.

[6] There is an additional technicality in this averaging, which is not of concern here. Because teachers with fewer students are more likely to get extreme value-added estimates just by chance, some models adjust for the amount of information available about each teacher using so-called shrinkage estimators to make extreme scores less likely. This is another modeling decision that influences the outcomes. Different models give different answers.

Briggs and Domingue (2011) reanalyzed the data used to generate the teacher effectiveness estimates published by the *Los Angeles Times* in August of 2010. Here is what they said about the statistical model used in the analyses published by the *LA Times*:

> The term "value-added" … is intended to have the same meaning as the term "causal effect" — that is, to speak of estimating the value-added by a teacher is to speak of estimating the causal effect of that teacher. But once stripped of the Greek symbols and statistical jargon, what we have left is a remarkably simple model that we will refer to as the "LAVAM" (Los Angeles Value-Added Model). It is a model which, in essence, claims that once we take into account five pieces of information about a student, the student's assignment to any teacher in any grade and year can be regarded as occurring at random. If that claim is accurate, the remaining differences can be said to be the value added or subtracted by that particular teacher. (Briggs & Domingue, 2011, p. 4)

The 5 pieces of information in the LAVAM were test performance in the previous year, gender, English language proficiency, eligibility for Title I services, and whether the student began schooling in the LA Unified School District after kindergarten. In effect, the LAVAM relies on these 5 variables to account for all the systematic differences among the students assigned to different teachers. My point here is not that this particular model is a bad one because it only includes 5 variables, although Briggs and Domingue (2011) did show that teacher rankings changed substantially when an alternative model with some additional control variables was used. (They interpreted their findings as showing that the alternative model had less bias.) The key point here is to understand how VAMs work: They adjust for some set of student characteristics, and sometimes for certain classroom or school characteristics, and then assume that once those adjustments are made, student assignments to teachers are as good as random.

Stated a little differently, the goal for the VAM is to strip away just those student differences that are outside of the current teacher's control — those things the teacher should *not* be held accountable for, leaving just those student test score influences the teacher *is* able to control and therefore *should* be held accountable for. This is a sensitive business, and different, defensible choices can lead to substantial differences in teachers' value-added rankings.

Earlier I offered an analogy of teacher value-added estimation being like a testing process, in which the teachers are the examinees and the classrooms full of students are like collections of items on different forms of a test. Before leaving that analogy, let me also point out that in any testing situation, common notions of fairness require that all examinees take the test under the same testing conditions. Unlike standardized testing conditions, in the VAM scenario the teacher-examinees may be working under far from equal conditions as they complete their value-added tests by teaching their students for a year. School climate and resources, teacher peer support, and, of course, the additional instructional support and encouragement students receive both out of school and from other school staff all make the test of teaching much easier for teachers in some schools and harder in others.

Statistical Assumptions

VAMs are complicated, but not nearly so complicated as the reality they are intended to represent. Any feasible VAM must rely on simplifying assumptions, and violations of these assumptions may increase bias or reduce precision of the model's value-added estimates. Violations of model assumptions also make it more difficult to quantify just how accurate or inaccurate those estimates really are. Hence, these statistical assumptions matter.

Effects of Social Stratification

Recall that the fundamental challenge is to estimate the average of each student's potential scores across all possible teachers. This is difficult due in part to the socioeconomic stratification in the U.S. school system. Reardon and Raudenbush (2009) pointed out that, "given the reality of school segregation on the basis of various demographic characteristics of students, including family socioeconomic background, ethnicity, linguistic background, and prior achievement … in practice, some students [may] have no access to certain schools" (p. 494). If teachers in some schools have virtually no access to high-achieving students from affluent families, and teachers in other schools have similarly limited access to low-achieving students from poor families, then the statistical model is forced to project well beyond the available data in order to estimate potential scores on a common scale for each student with each teacher. For this reason, VAM estimates are least trustworthy when they are used to compare teachers working in very different schools or with very different student populations.

Peer Effects

Another key assumption holds that a given student's outcome with a given teacher does not depend upon which other students are assigned to that same teacher. This is sometimes stated as *no peer effects*.[7] One usually thinks about *peer effects* as arising when students interact with each other. There are peer effects when small groups of students work collaboratively, for example. Or, peer effects are thought of as arising through peer culture — whether students reinforce or discourage one another's academic efforts. These kinds of effects are important, of course, but for value-added modeling, there are two *additional* kinds of peer effects that may be equally or more important.

The first of these has to do with how the members of the class *collectively* influence the teacher's pacing of instruction, the level at which explanations are pitched, the amount of reading assigned, and so forth. If the teacher is meeting the students where they are, then the average achievement level in the class as a whole is going to influence the amount of content delivered to *all* of the students over the course of the school year. In the real world of schooling, students are sorted by background and achievement through patterns of residential segregation, and they may also be grouped or tracked within schools. Ignoring this fact is likely to result in penalizing teachers of low-performing students and favoring teachers of high-performing students, just because the teachers of low-performing students cannot go as fast.

[7] Technically, *no peer effects* is an implication of the stable unit treatment value assumption (SUTVA).

Yet another kind of peer effect arises when some students in the classroom directly promote or disrupt the learning of others. Just about every teacher can recall some classes where the chemistry was right — perhaps one or two strong students always seemed to ask just the right question at just the right time to move the classroom discussion along. Most teachers can also recall some classes where things did not go so well. Perhaps one or two students were highly disruptive or repeatedly pulled the classroom discussion off topic, wasting precious minutes before the teacher could get the lesson back on track.[8] Simply put, the net result of these peer effects is that VAMs will *not* simply reward or penalize teachers according to how well or poorly they teach. The will *also* reward or penalize teachers according to *which students* they teach and *which schools* they teach in. Some of these peer effects (e.g., disruptive students) may add random noise to VAM estimates. Others (e.g., effect of average achievement level on pacing) may introduce bias.[9] Adjusting for individual students' prior test scores and other background characteristics may mitigate — but cannot eliminate — this problem.

[8] It is, of course, the teacher's responsibility to manage disruptive students, but the fact remains that teacher time spent dealing with such classroom disruptions may affect the learning of all students in the classroom.

[9] Some, but not all, VAMs incorporate classroom- or school-level measures to help control for these kinds of systematic effects.

An Interpretive Argument for Value-Added Model (VAM) Teacher Effectiveness Estimates

I suggested earlier that one might think of teacher value-added effectiveness estimates as a complicated kind of test score. Teachers are the examinees; each student is like a test item. Assigning a classroom full of students to a teacher for a year is like giving the teacher a test composed of 30 or so items. Thinking about the entire series of steps involved in value-added estimation as a single, complicated measurement process, one can consider the validity of VAM scores for any given purpose in much the same way as a testing expert would consider the validity of any other score. An *interpretive argument* is needed — a logical sequence of propositions that, taken together, make the case for the proposed use or interpretation. Then, once there is an interpretive argument, the strength of the evidence supporting each proposition must be considered.

Perhaps the most authoritative contemporary treatment of test validation is provided by Michael Kane's (2006) chapter, "Validation," in the most recent edition of *Educational Measurement*. His analysis laid out four broad steps in the interpretive argument (Kane, 2006, p. 34). My application of this framework to teacher VAM score estimation is shown in Table 2.

The first step is *scoring*. Here the scoring proposition holds that teacher VAM scores accurately capture each teacher's effectiveness, with the particular group of students that teacher actually taught, as measured by the student achievement test actually administered. In other words, each teacher's VAM score captures that teacher's degree of success in imparting the knowledge and skills measured by the student achievement test, reasonably undistorted by irrelevant factors. Scoring is the step from the teacher's classroom performance to the teacher's VAM score.

The second step is *generalization*, which addresses test reliability. One needs to know how stable VAM scores would be across different possible classes a

Table 2
An Interpretive Argument for Teacher Value-Added Model (VAM) Scores

Stage of interpretive argument	Description	Focusing question
1. Scoring *Observed score*	Construction of observed VAM score for an individual teacher	Is the score unbiased? (i.e., is systematic error acceptably small?)
2. Generalization *Observed score to universe score*	Generalization to scores that might have been obtained with a different group of students or a parallel form of the same test	Is the score reliable? (i.e., is random error acceptably small?)
3. Extrapolation *Universe score to target score*	Extrapolation to teacher effectiveness more broadly construed	Do scores correlate with other kinds of indicators of teaching quality? Do teacher rankings depend heavily on the particular test used? Does achievement test content fully capture valued learning outcomes? How do VAM scores relate to valued nontest (noncognitive) outcomes?
4. Implication *Target score to interpretation or decision*	Soundness of the intended decision or interpretation	Are intended benefits likely to be realized? Have plausible unintended consequences been considered?

teacher might have taught and also over time. If this year's VAM score gives poor guidance as to a teacher's likely effectiveness next year, then it is not very useful. In the language of test theory, this is the step from the observed score to the universe score — the long-run average across imagined repeated measurements.

The third step is *extrapolation*, which directs attention to the relation between the student achievement test actually used and other tests that might have been used instead for capturing student learning outcomes. It also covers the broader question of how well students' scores on this test *or similar tests* can capture the full range of important schooling outcomes. The real target of any measurement is some quality that is broader than test taking per se. In Kane's (2006) terminology, extrapolation is the move from the universe score to that target score.

Finally, the fourth step is *implication*, which directs attention to rationales for the expected benefits of each particular score use or interpretation, as well as plausible unintended consequences. This is the step from the target score to some decision or verbal description.

Let us next turn to some of the evidence concerning each of these steps. *Scoring* will address issues of *bias*, or *systematic* error. *Generalization* will address *reliability*, or *random* error. *Extrapolation* will address the relation between teacher VAM scores and *other measures* of effectiveness.[10] *Implication*, finally, will take up the question of appropriate and inappropriate *uses* of VAM scores and their likely *consequences*.

Scoring

Recall that the scoring step holds that a teacher's VAM estimate really does tell how effective that teacher was, this year, with these students, in teaching the content measured by this particular achievement test. This means the scoring must be free of systematic bias, the statistical model must reflect reality, and the data must fit the model. The word bias is used in a statistical sense, although here the commonsense meaning of the term is not too far off. Bias refers to errors that do not average out as more information is collected. If teachers in some kinds of schools, or working with some kinds of students, or teaching in certain grades or subject areas tend to get systematically lower or higher VAM estimates, that kind of error will not average out in the long run. The error will tend to show up again and again for a given teacher, in the same direction, year after year, simply because teachers tend to work with similar students year after year, typically in the same or similar schools.

Let us consider this question of bias. Jesse Rothstein (2010) published an important paper in which he developed and applied a *falsification test* for each of three different VAM specifications. Rothstein argued that it is logically impossible for current teacher assignments to influence students' test score gains in earlier years. This year's teacher cannot influence last year's achievement. Therefore, if a VAM is run backward in time, using current teacher assignments to predict students' score gains in *earlier* years, it ought to show that the true variance of prior year teacher effects, discounting random error, is near zero. This is called a *falsification test* because if the analysis does estimate substantial variance for prior

[10] Another important dimension of extrapolation is related to the assumption that a teacher's effectiveness with one sort of students is predictive of that teacher's effectiveness with different sorts of students. The assumption that a teacher has some effectiveness independent of the kinds of students that teacher is working with is important, but largely unexamined.

year teacher effects, then those estimates have to be biased. Such a finding strongly suggests that current-year teacher effect estimates may also be biased, although it does not prove the existence of bias.[11]

Rothstein (2010) tried this out using data from fifth grade classrooms in North Carolina. His sample included more than 60,000 students in more than 3,000 classrooms in 868 schools. He tried several different VAMs and consistently found that fifth grade teacher assignments showed powerful effects on third to fourth grade test score gains. Briggs and Domingue (2011) used Rothstein's test to look at the data on teachers from the LA Unified School District — the same data set Richard Buddin used to estimate the first round of teacher value-added scores published by the *Los Angeles Times* in August 2010. On the reading test, they found that teachers' estimated effects on their students' gains during a *previous* school year were about as large as their estimated effects on score gains during the current year. On a mathematics test, the logically impossible prior year effects came out around two thirds as large as for the current year. In one comparison, the estimated effects of fourth grade teachers on third grade reading gains were slightly *larger* than those teachers' estimated effects on fourth grade reading gains. Similar findings have emerged in other studies.

How can this be? As stated earlier, one reason is the massively nonrandom grouping of students, both within and between schools, as a function of family socioeconomic background and other factors. This clearly has the potential to distort teacher effectiveness estimates coming out of VAMs. Nonrandom assignment might also take the form of assigning struggling readers to reading specialists or English learners to bilingual teachers.

Bias is also possible due to differences in the schools where teachers work. Not all schools are equally conducive to student learning. Bias may come about because peer effects are not fully accounted for. Some limited evidence suggests that bias in VAMs may not be a serious problem (e.g., Chetty et al., 2011; Kane, McCaffrey, Miller, & Staiger, 2013). However, like all studies, each of these has some weaknesses and limitations.[12] Moreover, the fact that no bias is detected in one VAM application is no guarantee that bias may not exist in some other setting.

Another significant concern arises because the student achievement tests often used to date have been those mandated by NCLB (2002), which by law are limited to testing content at grade level. That means that teachers of gifted and talented classes may be unable to earn high value-added scores because their above grade level students are topping out on the tests and simply cannot demonstrate any further score gains. Likewise, teachers whose students are far below grade level may be penalized because the content they are teaching to meet their students' needs does not show up on the tests used to measure student growth.

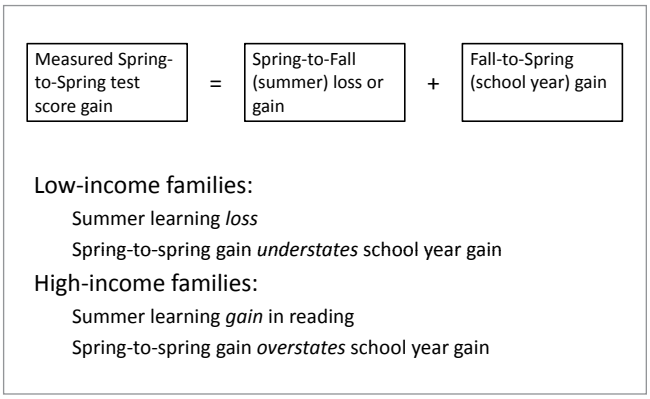
Yet another potential source of bias is related to *summer learning loss* (see Figure 3). Jennifer Sloan Mc-

[11] Goldhaber and Chaplin (2012) analyzed the conditions under which it is possible for one of Rothstein's specifications to yield a non-null finding even if current-year effect estimates are unbiased and called for further investigation. Chetty et al. (2011) implemented a quasi-experimental test for selection on unobservables, based on teacher switching between schools, and also concluded that, although they replicated Rothstein's results, this does not in fact imply that their estimates of long-term teacher effects are biased.

[12] The Chetty et al. (2011) study relied on student test data collected under relatively low-stakes conditions, which limits its applicability to VAMs with high stakes for teachers. The MET Project randomization study by Kane et al. (2013) examined random student assignment under rather constrained conditions and also suffered from problems of attrition and noncompliance. These problems limited its power to detect bias due to student assignment.

Combs and her colleagues at the RAND Corporation (McCombs et al., 2011) recently reviewed the research on summer learning loss. They concluded that on average, elementary school students lose about 1 month of learning over the summer months, from spring to fall. Losses are somewhat larger for mathematics, somewhat smaller for reading. But more importantly, these losses are not the same for all students. On average, students from higher income families actually post *gains* in reading achievement over the summer months, while their peers from lower income families post *losses*. This suggests a potential distortion in comparisons of VAM estimates among teachers whose students come from different economic backgrounds. On average, reading scores from the previous spring will *underestimate* the initial autumn proficiency of students in more advantaged classrooms and *overestimate* the initial autumn proficiency of those in less advantaged classrooms. Even if the two groups of students in fact make equal *fall*-to-spring gains, their measured prior *spring*-to-spring gains may differ. Some of this difference may be accounted for in VAMs that include adjustments for demographic factors, but once again, it appears likely that value-added estimates may be biased in favor of some teachers and against others.

Figure 3
Summer Learning Loss Is Not the Same for Students From Less Affluent Versus More Affluent Families



These concerns must be balanced against compelling empirical evidence that teacher VAM scores are capturing some important elements of teaching quality. In particular, Chetty et al. (2011) recently reported that teachers' VAM scores predicted their students' future college attendance, earnings, socioeconomic status, and even teenage pregnancy rates.[13] Their study included creative statistical tests for bias due to omitted variables, and they found no bias. Similarly, Goldhaber and Hansen (2010) have reported modest but statistically significant effects of teacher VAM estimates on student test scores several years later. Teacher VAM scores are certainly not just random noise. These models appear to capture important differences in teachers' effects on student learning outcomes. But even the best models are not *pure* measures of teacher effectiveness. VAM scores *do* predict important student learning outcomes, but my reading of the evidence strongly suggests that these scores nonetheless measure not only *how well* teachers teach, but also *whom* and *where* they teach.

[13] The study by Chetty et al. (2011) is very carefully done, but relied on data collected in a context in which no particularly high stakes were attached to student test scores. Even in that context, the authors set aside the top 2% of teacher VAM scores because "these teachers' impacts on test scores appear suspiciously consistent with testing irregularities indicative of cheating" (Chetty et al., 2011, p. 23). When these teachers were included in the analysis, estimated long-term teacher effects were reduced by roughly 20% to 40%.

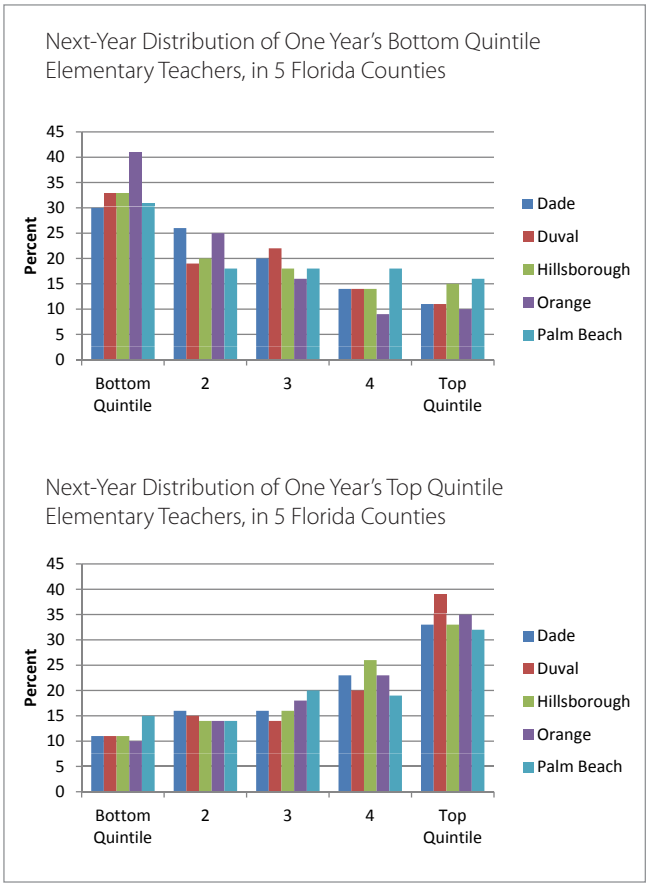
Generalization

The second link in the chain of propositions needed to support VAM scores is generalization, the step from observed score to universe score. The first proposition, scoring, focused on the question of *what* value-added scores were measuring, including the question of whether those scores were free from systematic bias. Generalization shifts attention from *what* to *how well* and from *systematic* error to *random* error. It focuses on the question of how *stable* or *unstable* teacher VAM scores turn out to be. This is the familiar issue of score reliability.

One very good way to estimate reliability is just to correlate value-added scores from two points in time, or from two sections of the same class. The correlation itself is the same as a reliability coefficient. Several years ago, Daniel McCaffrey and his co-authors investigated a variety of VAM specifications and data sets and found year-to-year correlations mostly between .2 and .4, with a few lower and a few higher (McCaffrey, Sass, Lockwood, & Mihaly, 2009). More specifically, they looked at value-added scores for teachers in five different counties in Florida. Figure 4 illustrates some of their findings for elementary school teachers. They found that in each county, a minimum of 10% of the teachers in the bottom fifth of the distribution one year were in the top fifth the next year, and conversely. Typically, only about a third of 1 year's top performers were in the top category again the following year, and likewise, only about a third of 1 year's lowest performers were in the lowest category again the following year. These findings are typical. A few studies have found reliabilities around .5 or a little higher (e.g., Koedel & Betts, 2007), but this still says that only half the variation in these value-added estimates is signal, and the remainder is noise.

Figure 4

Year-to-Year Changes in Teacher Value-Added Rankings Reported by McCaffrey et al. (2009, Table 4, p. 591)



McCaffrey and his colleagues (2009) pointed out that year-to-year changes in teachers' scores reflected *both* the vagaries of student sampling *and* actual changes in teachers' effectiveness from year to year. But if one wants to know how useful one year's score is for predicting the next year's score, that distinction does not matter. McCaffrey et al.'s results imply that unstable or random components together account for more than half the variability in VAM scores, and in some cases as much as 80% or more. Sorting teachers according to single year value-added scores is sorting mostly on noise.

One of the most important ongoing studies of value-added modeling, classroom observation, and other methods of teacher evaluation is the Measures of Effective Teaching (MET) project sponsored by the Bill & Melinda Gates Foundation. VAM reliabilities reported by the MET project are consistent with earlier estimates from other studies. A 2010 MET report gives correlations between VAM scores for 2 successive years, as well as correlations between estimates obtained for two different sections of the same course, taught the same year to different students. As shown in Table 3, on 4 different tests, 2 in mathematics and 2 in English language arts, the correlations between sections taught the same year range from .18 to .38, and the correlations across years, available for just one of the mathematics tests and one of the reading tests, are .40 and .20, respectively (the MET Project, 2010, p. 18). These numbers are well under .5, which means that once again, in all cases, over half the variation in teachers' single year value-added estimates is random or unstable.

Table 3
Teacher VAM Reliabilities Reported From the MET Project

Test	Same year, different course selection	Different year
State mathematics test	0.381	0.404
State English language arts test	0.180	0.195
Balanced assessment in mathematics	0.228	
Stanford 9 open-ended reading	0.348	

Note: Data from the MET Project (2010).

On standardized tests with stakes for students, reliability coefficients of at least .80, preferably .85 or .90, are the goal. A coefficient of .80 means that 80% of the variation in scores is attributable to real differences in the attribute the test measures, and only 20% is measurement error. Value-added reliabilities of .2 to .5 imply that as little as 20% of the score variation is attributable to the quality the scores are measuring and as much as 80% is due to measurement error.

Of course, reliability of VAM scores can be increased considerably by pooling results over 2 or 3 years. If the reliability of single year VAM scores were .30, say, then a 2 year rolling average should have a reliability of roughly .46, and a 3 year rolling average, roughly .56 — these numbers are still not very good, but they are much improved over single year estimates. Unfortunately, many VAM implementations have relied on results from just a year at a time.

It seems clear from anecdotal accounts that teachers are troubled by the year-to-year fluctuations they see in value-added effectiveness estimates. In our paper last year in the *Phi Delta Kappan* (Darling-Hammond, Amrein-Beardsley, Haertel, & Rothstein, 2012), my colleagues and I included this quote from a teacher in Houston:

> I do what I do every year. I teach the way I teach every year. [My] first year got me pats on the back; [my] second year got me kicked in the backside. And for year three, my scores were off the charts. I got a huge bonus, and now I am in the top quartile of all the English teachers. What did I do differently? I have no clue. (Amrein-Beardsley & Collins, 2012, p. 15)

In Nashville, Tennessee, where teachers are well familiar with this form of teacher evaluation, almost 300 teachers were among the initial volunteers for a 3 year study of teacher merit pay based on value-added scores, called the POINT experiment. From the outset, almost

70% of the teachers *disagreed* with the statement: "The POINT experiment will do a good job of distinguishing effective from ineffective teachers in the treatment group" (Springer et al., 2010, p. 38).[14]

Extrapolation

Moving on from scoring and generalization, I come to the third proposition, *extrapolation*. This is the step from universe score to target score, asking how well value-added scores reflect broader notions of teacher effectiveness. Different kinds of evidence can be brought to bear here, framed as four questions. First, one can ask how well teacher VAM estimates track other kinds of information about teaching quality. Second, one can ask how much the estimates change if a different test is used. Third, one can consider whether the achievement tests typically used for these models reflect the range of desired cognitive outcomes. Finally, one can look beyond test scores to a broader range of learning outcomes, including noncognitive skills not directly reflected in test scores.[15]

How well do teacher VAM estimates track other kinds of information about teaching quality? Hill, Kapitula, and Umland (2010) reported on a study that delved deeply into teachers' instructional practices, with multiple in-depth interviews, classroom observations, and detailed analyses of lesson transcripts. This is a different kind of research from the large correlational studies. Only 24 teachers were involved, all teaching middle school mathematics. But Hill et al.'s findings are quite revealing. The authors present case studies of two teachers tied for the lowest possible rating on their scale for mathematical quality of instruction. Here is what they said about the first of the two:

> In every [lesson], there were significant problems with the basic mathematics of middle school. She reasons incorrectly about unit rates. She concludes that an answer of 0.28 minutes must actually be 0.28 seconds because one cannot have a fraction of a minute. She tells students that integers include fractions. She reads a problem out of the text as 3/8 +2/7 but then writes it on the board and solves it as 3.8 + 2.7. She calls the commutative property the community property. She says proportion when she means ratio. She talks about denominators being equivalent when she means the fractions are equivalent. (Hill et al., 2010, p. 820)

Hill et al. (2010) reported that this teacher's value-added score was in the second highest quartile.[16] They thoughtfully considered whether her instruction might have had other redeeming features that compensated for her basic lack of mathematical competence and were able to find none.

---

[14] These survey responses may reflect teachers' concerns over bias (systematic error) as well as reliability (random error).

[15] A further dimension of extrapolation, not considered in this paper, is the extent to which a teacher's effectiveness in the sort of school and with the sort of students observed is predictive of that teacher's effectiveness in different teaching contexts.

[16] Hill et al. (2010) took VAM scores for the 24 teachers observed from a larger set of VAM estimates for 222 teachers in the district. Eight different covariance-adjustment models were investigated, and findings were reported for a simple model (average gain scores) as well as a model adjusting for student demographic variables and a model adjusting for school fixed effects. The findings in this report use the simple model VAM estimates. Results might differ with a more sophisticated value-added model, although the authors reported that teacher rankings across the various models considered were all highly correlated.

Concerning the other bottom tier performer, the authors wrote that:

> The overwhelming impression of [his] classroom is that there is very little mathematics occurring. In many lessons, [he] offers only the briefest of mathematical presentations, typically referring students to the text and assigning a series of problems. In one lesson, he fails altogether to directly teach any material. And throughout the classes we observed, student behavior is a serious issue. (Hill et al., 2010, p. 820)

This teacher's mathematics value-added score was in the top quartile, presumably because he was teaching a classroom of accelerated students. These 2 teachers sound like the kinds of bad teachers one might imagine VAM scores should catch, but they both sailed right through.

Table 4
MET Project Correlations Between Value-Added Model (VAM) Scores and Classroom Observations

Subject area	Classroom observation system	Correlation of overall quality rating with prior year VAM score
Mathematics	CLASS	0.18
Mathematics	FFT	0.13
Mathematics	UTOP	0.27
Mathematics	MQI	0.09
English language arts	CLASS	0.08
English language arts	FFT	0.07
English language arts	PLATO	0.06

Note: Data are from the MET Project (2012, pp. 46, 53). CLASS = Classroom Assessment Scoring System, FFT = Framework for Teaching, PLATO = Protocol for Language Arts Teaching Observations, MQI = Mathematical Quality of Instruction, UTOP = UTeach Teacher Observation Protocol.

There is also some evidence from the MET Project relating VAM scores to other indicators of teaching quality. First, the MET Project (2012) provided correlations between overall instructional quality, as measured by each of 4 different classroom observation systems, and prior year teacher VAM scores. These correlations ranged from .06 to .27 (the MET Project, 2012, pp. 46, 53). Consistent with the study by Hill and her colleagues (Hill et al., 2010), these correlations imply positive, but very weak, observed relationships between VAM scores and teaching quality as indicated by direct classroom observations (see Table 4).

The MET Project (2010) also included classroom climate surveys completed by students themselves, with questions focused on specific aspects of teaching practice. These surveys were designed to show whether students experienced their classrooms as engaging, demanding, and supportive of intellectual growth. One might expect that teacher's VAM scores would track their students' perceptions, but as shown in Table 5, the correlations between VAM scores and overall classroom climate ratings ranged from only .06 to .22 (the MET Project, 2010, pp. 23, 25).

A second line of evidence about the extrapolation proposition comes from comparisons of VAM scores for the same students and teachers, but using different tests. J. R. Lockwood and colleagues (Lockwood et al., 2007) compared value-added estimates obtained using two different subtests of the same mathematics test. One subtest was on procedures, the other covered problem solving. Using different models, with different sets of control variables, Lockwood et al. obtained correlations ranging from .01, essentially a zero relationship, to a high of .46, with a median value of .26, between teachers' value-added scores based on one subtest versus the other. In the words of the authors, "These correlations are uniformly low … the two achievement outcomes

lead to distinctly different estimates of teacher effects." (Lockwood et al., 2007, p. 54)

Table 5
MET Project Correlations Between Value-Added Model (VAM) Scores and Student Classroom Climate Surveys

Test	Same year, same section	Same year, different section
State mathematics test	.21	.22
State English language arts test	.10	.07
Balanced assessment in mathematics	.11	.11
Stanford 9 open-ended reading	.14	.06

Note: Data are from the MET Project (2010, pp. 23, 25).

John Papay (2011) did a similar study using three different reading tests, with similar results. He stated his conclusion as follows:

> [T]he correlations between teacher value-added estimates derived from three separate reading tests — the state test, SRI [Scholastic Reading Inventory], and SAT [Stanford Achievement Test] — range from 0.15 to 0.58 across a wide range of model specifications. Although these correlations are moderately high, these assessments produce substantially different answers about individual teacher performance and do not rank individual teachers consistently. Even using the same test but varying the timing of the baseline and outcome measure introduces a great deal of instability to teacher rankings. Therefore, if a school district were to reward teachers for their performance, it would identify a quite different set of teachers as the best performers depending simply on the specific reading assessment used. (Papay, 2011, p. 187)

Once more, the MET study offered corroborating evidence. The correlation between value-added scores based on two different mathematics tests given to the same students the same year was only .38. For 2 different reading tests, the correlation was .22 (the MET Project, 2010, pp. 23, 25).

So the pattern is pretty consistent. Teacher VAM scores offer surprisingly little information even about the same teachers' scores with the same students, the same year, based on a different test. With regard to the second of the four questions in this section, value-added measures do not fare very well. Rankings change substantially, simply as a function of the reading or mathematics test chosen.

My third question in this section concerns the contents of the tests. A recent article by Polikoff, Porter, and Smithson (2011) summarized and updated studies by several researchers concerning alignment of state tests to academic content standards. These studies typically showed that state tests place too much emphasis on memorization and too little on complex cognitive processes. Here is what Polikoff and his colleagues concluded:

> Data from … 19 states were used here to investigate whether state standards and assessments under NCLB are in fact aligned ... Clearly, when alignment is defined in the way that is most predictive of value-added to student achievement … the answer is no. (Polikoff et al., 2011, p. 989)

Of course, the Smarter Balanced Assessment Consortium and the Partnership for Assessment of Readiness for College and Careers (PARCC) soon will launch new tests designed to reflect the new Common Core State Standards (National Governors Association & Council

of Chief State School Officers, 2010), and so the alignment picture may improve somewhat. But history suggests that teachers will still feel pressure to focus just on tested content, at the expense of important learning outcomes that are more difficult to measure (see, e.g., Koretz, 2008; Madaus, 1988).

The first 3 questions in this section asked about VAM scores versus other evidence regarding teaching quality, about VAM scores based on different tests, and about content coverage on state tests under NCLB. The final question here may be the most important of all, but little direct, empirical evidence can be brought to bear. That question addresses extrapolation beyond test scores toward a broader range of schooling outcomes, including noncognitive outcomes.

The profound importance of noncognitive skills for economic as well as educational outcomes has been clearly established by the work of James Heckman and his colleagues at the University of Chicago, among others (e.g., Borghans, Duckworth, Heckman, & ter Weel, 2008). A recent review by Henry M. Levin (2012) argued forcefully that our fixation on cognitive achievement tests to measure schooling success may be misguided. Levin is both an educational researcher and an economist. He reviewed findings about human capital formation and concluded not only that noncognitive skills are important, but that they can be developed deliberately through education.

Acknowledging that student achievement tests do not measure noncognitive outcomes directly, proponents of value-added modeling still have a strong argument that those teachers most effective in raising test scores may also be those most effective in fostering other kinds of learning. The empirical evidence from the Chetty et al. (2011) study would seem to show that *whatever* it is test scores are picking up, it matters. (Note, however, that their findings depended on test scores obtained under low-stakes conditions.) Moreover, many educators are concerned that these noncognitive skills, along with non-tested subjects such as art, music, or even science, may be driven out of the curriculum as educators respond to more and more powerful pressures, such as test-score based teacher evaluation, to teach just to the high-stakes tests. At the outset of the Nashville POINT experiment, mentioned earlier, 80% of the teachers *agreed* with the statement: "The POINT experiment ignores important aspects of my performance that are not measured by test scores." Two years later, that figure was essentially unchanged, at 85% (Springer et al., 2010).

Implication

The final step in the interpretive argument, *implication*, moves from target score to verbal description. With this step, it becomes even more important to be specific about the details of how scores will be used. The same test can be valid for one purpose and invalid for another, and no test is valid for all purposes. As stated at the outset, this discussion of scoring, generalization, and extrapolation has focused primarily on high-stakes uses with consequences for individual teachers. The evidence pro and con could be weighted differently if other uses were considered.

My first conclusion should come as no surprise: Teacher VAM scores should emphatically *not* be included as a substantial factor with a fixed weight in consequential teacher personnel decisions. The information they provide is simply not good enough to use in that way. It is not just that the information is noisy. Much more serious is the fact that the scores may be systematically biased *for* some teachers and *against* others,

and major potential sources of bias stem from the way our school system is organized. No statistical manipulation can assure fair comparisons of teachers working in very different schools, with very different students, under very different conditions. One cannot do a good enough job of isolating the signal of teacher effects from the massive influences of students' individual aptitudes, prior educational histories, out-of-school experiences, peer influences, and differential summer learning loss, nor can one adequately adjust away the varying academic climates of different schools. Even if acceptably small bias from all these factors *could* be assured, the resulting scores would still be highly unreliable and overly sensitive to the particular achievement test employed. Some of these concerns can be addressed, by using teacher scores averaged across several years of data, for example. But the interpretive argument is a *chain* of reasoning, and every proposition in the chain must be supported. Fixing one problem or another is not enough to make the case.

If there were sound evidence that value-added teacher evaluation improves student learning, one might still try to make them work for this purpose. There is certainly solid evidence that VAMs can detect real differences in teacher effectiveness, but as noted, some of the strongest evidence (e.g., Chetty et al., 2011; the MET Project, 2010, 2012) has come from studies in which student test scores were not high stakes for teachers.[17] Students certainly benefit from more effective teachers, and teacher effectiveness may matter most for students from disadvantaged backgrounds. But there are substantial obstacles to translating these substantive findings into successful educational policies. High-stakes uses of teacher VAM scores could easily have additional *negative* consequences for children's education. These include increased pressure to teach to the test, more competition and less cooperation among the teachers within a school, and resentment or avoidance of students who do not score well. In the most successful schools, teachers work together effectively (Atteberry & Bryk, 2010). If teachers are placed in competition with one another for bonuses or even future employment, their collaborative arrangements for the benefit of individual students as well as the supportive peer and mentoring relationships that help beginning teachers learn to teach better may suffer.

Are teacher-level VAM scores good for anything, then? Yes, absolutely. But, for some purposes, they must be used with considerable caution. To take perhaps the easiest case first, for researchers comparing large groups of teachers to investigate the effects of teacher training approaches or educational policies, or simply to investigate the size and importance of long-term teacher effects, it is clear that value-added scores are far superior to unadjusted end-of-year student test scores.[18] Averaging value-added scores across many teachers will damp down the random noise in these estimates and could also help with some of the systematic biases, although that is not guaranteed. So, for research purposes, VAM estimates definitely have a place. This is also one of the safest applications of VAM scores because the policy researchers applying these models are likely to have the training and expertise to respect their limitations.

[17] In the MET Project (2010, p. 21), value-added scores from high-stakes state tests were not strongly predictive of scores from specially administered lower stakes tests.

[18] Note that my reference here to *teacher training approaches* is intended to refer to broad research questions comparing groups of programs that share common features. An example of this kind of research is the study by Boyd, Grossman, Lankford, Loeb, and Wyckoff (2009). I would emphatically *not* advocate the use of teacher value-added scores to evaluate or compare individual teacher preparation programs.

A considerably riskier use, but one I would cautiously endorse, would be providing individual teachers' VAM estimates to the teachers themselves and to their principals, provided *all 5* of the following critically important conditions are met:

- Scores based on sound, appropriate student tests
- Comparisons limited to homogeneous teacher groups
- No fixed weight — flexibility to interpret VAM scores in context for each individual case
- Users well trained to interpret scores
- Clear and accurate information about uncertainty (e.g., margin of error)

First, the scores must be based on sound and appropriate student achievement tests, aligned to the content teachers are expected to cover, providing valid measurements for the full range of student achievement levels, and scaled appropriately. This may sound obvious, but it is in fact a very strong limitation on the applicability of these models. One of the most troubling aspects of some current reform proposals is the insistence on *universal* application of value-added to all teachers in a district or state. For most teachers, appropriate test data are not available, period. They teach children so young that there are no prior year scores, or they teach untested subjects, or they teach high school courses for which there are no pretest scores that it makes any sense to use.

Second, comparisons should be limited to fairly homogeneous groups of teachers. Rankings that mix teachers from different grade levels or teaching in schools with very different demographics place severe demands on statistical model assumptions, and the effects of violations of these assumptions are often not well understood. Conservative, local comparisons restricted to single subject areas and grade levels within homogeneous districts are much safer.

Third, there should be *no fixed weight* attached to the scores in reaching any consequential decisions. Principals and teachers must have the latitude to set aside an individual's score entirely — to ignore it completely — if they have specific information about the local context that could plausibly render that score invalid.

Fourth, anyone using teacher VAM scores in consequential decisions must be well trained to interpret the scores appropriately.

And, finally, score reports must be accompanied by clear and comprehensible information about the scores' instability and imprecision, and the range of factors that could render the scores invalid.

These 5 conditions would be tough to meet, but regardless of the challenge, if teacher value-added scores cannot be shown to be valid for a given purpose, then they should not be used for that purpose.

So, in conclusion, VAMs may have a modest place in teacher evaluation systems, but only as an adjunct to other information, used in a context where teachers and principals have genuine autonomy in their decisions about using and interpreting teacher effectiveness estimates in local contexts.

Sound Teacher Evaluation

What is there besides value-added? Better teacher evaluation methods share several common features. First, they attend to what teachers actually do — someone with training looks directly at classroom practice or at records of classroom practice such as teaching portfolios. Second, they are grounded in the substantial research literature, refined over decades of research, that specifies effective teaching practices. I am certainly not suggesting that there is just one way to teach, but there are some commonalities. Good teachers know their subject matter well and understand the kinds of misconceptions students often have and how to help them. They connect new learning to prior learning and, where appropriate, they build on students' own out-of-school experiences. They monitor and assess frequently, set clear standards, and provide feedback, for example. Third, because sound teacher evaluation systems examine what teachers actually do in the light of best practices, they provide constructive feedback to enable improvement. This means that these evaluation approaches can both guide teacher improvement and support timely and efficient personnel decisions.

Please note that I am *not* saying here, "Just do classroom observations instead of value-added." Classroom observation per se is not magic. Long experience has shown that classroom observation can be done very badly. Observations need to be systematic, observers need to be well qualified and well trained, and systems have to be in place for observers to document what they see and to monitor observer reliability. Observations must be followed up with feedback. Moreover, classroom observations are generally no more reliable than VAM scores, and they are doubtless subject to some of the same biases as VAM scores are. From my own work years ago on performance-based measures of teaching, I know that there are classrooms full of students that can make just about any teacher look good and others that will challenge the best teachers just to maintain a classroom environment conducive to learning.

There have been proposals to use VAM scores as a sort of trigger for more careful follow-up evaluation. This might be a reasonable stopgap approach if carefully implemented. It certainly makes sense for principals to spend most of their time focusing on the teachers who need help. However, the simple trigger idea still leaves much to be desired. First, as noted, VAM scores are only going to be available for a minority of teachers in most school systems, unless we undertake a vast expansion of student testing. Second, many weak teachers in need of assistance will be missed simply because their VAM scores will look just fine. Finally, this approach turns teacher evaluation into a sort of remediation strategy. I would prefer to regard sound professional evaluation and opportunities for continuous improvement as something that *all* teachers ought to be entitled to.

References

Amrein-Beardsley, A., & Collins, C. (2012). The SAS Education Value-Added Assessment System (SAS® EVAAS®) in the Houston Independent School District (HISD): Intended and unintended consequences. *Education Policy Analysis Archives*, *20*(12). Retrieved from **http://epaa.asu.edu/ojs/article/view/1096**

Atteberry, A., & Bryk, A. S. (2010). Centrality, connection, and commitment: The role of social networks in a school-based literacy initiative. In A. J. Daly (Ed.), *Social network theory and educational change* (pp. 51–75). Cambridge, MA: Harvard University Press.

Borghans, L., Duckworth, A. L., Heckman, J. J., & ter Weel, B. (2008). The economics and psychology of personality traits. *Journal of Human Resources*, *43*(4), 972–1059.

Boyd, D. J., Grossman, P. L., Lankford, H., Loeb, S., & Wyckoff, J. (2009). Teacher preparation and student achievement. *Educational Evaluation and Policy Analysis*, *31*(4), 416–440.

Briggs, D., & Domingue, B. (2011). *Due diligence and the evaluation of teachers: A review of the value-added analysis underlying the effectiveness rankings of Los Angeles Unified School District teachers by the* Los Angeles Times. Boulder, CO: National Education Policy Center. Retrieved from **http://nepc.colorado.edu/publication/due-diligence**

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2011). *The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood* (NBER Working Paper 17699). Retrieved from **http://www.nber.org/papers/w17699**

Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E., & Rothstein, J. (2012). Evaluating teacher evaluation. *Phi Delta Kappan*, *93*(6), 8–15.

An education exchange with Linda Darling-Hammond. (2011, April 4). *UCLA/IDEA Newsroom.* Retrieved from **http://idea.gseis.ucla.edu/newsroom/idea-news/an-education-exchange-with-linda-darling-hammond**

Goldhaber, D. D., Brewer, D. J., & Anderson, D. J. (1999). A three-way error components analysis of educational productivity. *Education Economics*, *7*(3), 199–208.

Goldhaber, D., & Chaplin, D. (2012). *Assessing the "Rothstein test": Does it really show teacher value-added models are biased?* (National Center for Analysis of Longitudinal Data in Education Research [CALDER] Working Paper 71). Washington, DC: The CALDER Center. Retrieved from **http://www.caldercenter.org/publications/calder-working-paper-71.cfm**

Goldhaber, D., & Hansen, M. (2010). *Assessing the potential of using value-added estimates of teacher job performance for making tenure decisions* (National Center for Analysis of Longitudinal Data in Education Research [CALDER] Working Paper 31). Washington, DC: The Urban Institute. Retrieved from **http://www.urban.org/publications/1001369.html**

Hill, H. C., Kapitula, L., & Umland, K. (2010). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal*, *48*(3), 794–831.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education/Praeger.

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). *Have we identified effective teachers? Validating measures of effective teaching using random assignment* (MET Project Research Paper). Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from **http://www.metproject.org/reports.php**

Koedel, C., & Betts, J. R. (2007). *Re-examining the role of teacher quality in the educational production function* (University of Missouri-Columbia Department of Economics Working Paper Series WP 07-08). Retrieved from **http://econ.missouri.edu/working-papers/2007/wp0708_koedel.pdf**

Koretz, D. (2008). *Measuring up: what educational testing really tells us*. Cambridge, MA: Harvard University Press.

Kupermintz, H., Shepard, L., & Linn, R. (2001, April). *Teacher effects as a measure of teacher effectiveness: construct validity considerations in TVAAS (Tennessee Value Added Assessment System)*. Paper presented at the meeting of the National Council on Measurement in Education, Seattle, Washington.

Levin, H. M. (2012). More than just test scores. *Prospects*, *42*(3), 269–284.

Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V., & Martinez, J. F. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement*, *44*(1), 47–67.

Madaus, G. F. (1988). The distortion of teaching and testing: high-stakes testing and instruction. *Peabody Journal of Education*, *65*(3), 29–46.

McCaffrey, D. F., Sass, T. R., Lockwood, J. R., & Mihaly, K. (2009). The intertemporal variability of teacher effect estimates. *Education Finance and Policy*, *4*(4), 572–606.

McCombs, J. S., Augustine, C. H., Schwartz, H. L., Bodilly, S. J., McInnis, B., Lichter, D. S., & Cross, A. B. (2011). *Making summer count. How summer programs can boost children's learning*. Santa Monica, CA: The RAND Corporation. Retrieved from **http://www.rand.org/content/dam/rand/pubs/monographs/2011/RAND_MG1120.pdf**

The MET Project. (2010). *Learning about teaching: initial findings from the Measures of Effective Teaching Project* (MET Project Research Paper). Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from **http://www.metproject.org/downloads/Preliminary_Findings-Research_Paper.pdf**

The MET Project. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains* (MET Project Research Paper). Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from **http://www.metproject.org/downloads/MET_Gathering_Feedback_Research_Paper.pdf**

National Governors Association, & Council of Chief State School Officers. (2010). *Common Core State Standards*. Washington, DC: Author.

No Child Left Behind Act of 2001, 20 U.S.C. § 6301 *et seq.* (2002).

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis*, *26*(3), 237–257.

Papay, J. P. (2011). Different tests, different answers: the stability of teacher value-added estimates across outcome measures. *American Educational Research Journal*, *48*(1), 163–193.

Polikoff, M. S., Porter, A. C., & Smithson, J. (2011). How well aligned are state assessments of student achievement with state content standards? *American Educational Research Journal*, *48*(4), 965–995.

Ravitch, D. (2010). *The death and life of the great American school system*. Philadelphia, PA: Basic Books.

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

Reardon, S. F., & Raudenbush, S. W. (2009). Assumptions of value-added models for estimating school effects. *Education Finance and Policy*, *4*(4), 492–519.

Sanders, W. L., & Rivers, J. C. (1996, November). *Cumulative and residual effects of teachers on future student academic achievement* (Research Progress Report). Knoxville, TN: University of Tennessee Value-Added Research and Assessment Center.

Springer, M. G., Ballou, D., Hamilton, L., Le, V., Lockwood, J. R., McCaffrey, D., Pepper, M., … Stecher, B. (2010). *Teacher pay for performance: Experimental evidence from the project on incentives in teaching*. Nashville, TN: National Center on Performance Incentives at Vanderbilt University. Retrieved from **http://www.rand.org/content/dam/rand/pubs/reprints/2010/RAND_RP1416.pdf**

Winerip, M. (2011, November 6). In Tennessee, following the rules for evaluations off a cliff. *The New York Times*. Retrieved from **http://www.nytimes.com/2011/11/07/education/tennessees-rules-on-teacher-evaluations-bring-frustration.html**

About ETS

At ETS, we advance quality and equity in education for people worldwide by creating assessments based on rigorous research. ETS serves individuals, educational institutions and government agencies by providing customized solutions for teacher certification, English language learning, and elementary, secondary and post-secondary education, as well as conducting education research, analysis and policy studies. Founded as a nonprofit in 1947, ETS develops, administers and scores more than 50 million tests annually — including the *TOEFL*® and *TOEIC*® tests, the *GRE*® tests and *The Praxis Series*™ assessments — in more than 180 countries, at over 9,000 locations worldwide.

23497

Listening. Learning. Leading.®

www.ets.org