ETS®

Policy Information **PERSPECTIVE**

Using Student Progress To Evaluate Teachers: A Primer on Value-Added Models

by Henry I. Braun

Policy
Information
Center

Listening. Learning. Leading.

Additional copies of this report can be
ordered for $10.50 each (prepaid) from:
Policy Information Center
Mail Stop 19-R
Educational Testing Service
Rosedale Road
Princeton, NJ 08541-0001
(609) 734-5949
e-mail: pic@ets.org
Copies can be downloaded for free at:
www.ets.org/research/pic

September 2005

Table of Contents

Preface

The concept is simple and attractive: Evaluate teachers on the basis of how much academic growth their students experience over the course of the school year. Often referred to as "value-added," this concept and the statistical methods for implementing it have been a topic of debate in state legislatures and at state and national education conferences over the past decade.

Recently, the concept and the practice have also been catching on in schools, districts and states across the country. Results from value-added models are already playing an increasing role in the process of identifying teachers in need of targeted professional development. But, as is often the case, the issues involved in implementing this seemingly straightforward idea are complex and pose both statistical and practical challenges.

In this Policy Information Perspective, Henry Braun examines value-added models and concludes with advice for policymakers who are seeking to understand both the potential and the limitations inherent in using such models to evaluate teachers. While welcoming the possibility of introducing a quantitative component into the teacher evaluation process, Henry Braun counsels policymakers to move forward with caution, especially if high stakes are attached to the results. ■

Michael T. Nettles
Vice President
Policy Evaluation & Research Center

Acknowledgments

Patty McAllister first suggested the need for such a primer on the use of value-added models to evaluate teachers. She and Penny Engel were instrumental in providing both support and advice in the early stages of this effort. The following individuals offered many useful comments on various drafts of the manuscript: Dale Ballou, Paul Barton, Beth Ann Bryant, Richard Coley, Carol Dwyer, Daniel Eignor, Emerson Elliott, Howard Everson, Sandy Kress and Robert Linn. The author also benefited from conversations with Harold Doran, Les Francis, Ted Hershberg, Irwin Kirsch, Daniel McCaffrey, William Sanders, Howard Wainer and Michael Zieky. Errors of fact or interpretation, however, are those of the author. Amanda McBride was the editor, and Marita Gray designed the cover. ■

Executive Summary

The quantitative evaluation of teachers based on an analysis of the test score gains of their students is an exciting prospect that has gained many proponents in recent years. Such evaluations employ a class of statistical procedures called "value-added models" (VAMs). These models require data that track individual students' academic growth over several years and different subjects in order to estimate the contributions that teachers make to that growth. Despite the enthusiasm these models have generated among many policymakers, several technical reviews of VAMs have revealed a number of serious concerns. Indeed, the implementation of such models and the proposed uses of the results raise a host of practical, technical, and even philosophical issues.

This report is intended to serve as a layperson's guide to those issues, aiding interested parties in their deliberations on the appropriate uses of a powerful statistical tool. It counsels caution and the need to carry out due diligence before enshrining such procedures into law. Although this report pays special attention to the VAM developed by William Sanders — which is now used by districts in such states as Tennessee, Ohio and Pennsylvania — much of the discussion applies to all VAMs.

First and foremost, treating the output of a value-added analysis as an accurate indicator of a teacher's relative contribution to student learning is equivalent to making a causal interpretation of a statistical estimate. Such interpretations are most credible when students are randomly sorted into classes, and teachers are randomly assigned to those classes. In the absence of randomization, causal interpretations can be misleading.

In reality, the classroom placement of students and teachers is far from random. In most districts, parents often influence where their children go to school and even to which class and teacher they are assigned. Similarly, teachers may select the school and classroom where they are placed.

Thus, the students assigned to a particular teacher may not be representative of the general student population with respect to their level and rate of growth in achievement, parental support, motivation, study habits, interpersonal dynamics and other relevant characteristics. It is very difficult for the statistical machinery to disentangle these intrinsic student differences from true differences in teacher effectiveness.

Student progress can also be influenced by the physical condition of the school and the resources available, as well as school policies and school-level implementation of district policies — all of which are beyond a teacher's control. To the extent that these characteristics vary systematically across schools in the district, they can undermine the fairness of a value-added approach to teacher evaluation.

Other issues discussed in this report include the nature of the test scores that serve as the raw material for VAMs, the amount of information available to estimate each teacher's effectiveness, and the treatment of missing data, which is endemic in district databases. Fortunately, a great deal of research is being undertaken to address each of these issues, and the report provides many relevant references. New studies of different VAMs, in a variety of settings, are providing a clearer picture of the strengths and limitations of the various approaches.

Notwithstanding the report's emphasis on caution, the widespread interest in VAMs should be welcomed. It has helped to move the conversation about teacher quality to where it belongs — on increasing student learning as the primary goal of teaching. It also introduces the promise of a much-needed quantitative component in teacher evaluation, while prompting a reexamination of issues of fairness and proper test use. These are steps in the right direction. By relying on measures of student growth, VAMs may ultimately offer a more defensible foundation for teacher evaluation than, say, methods based on absolute levels of student attainment or the proportion of students meeting a fixed standard of performance.

Given their current state of development, VAMs can be used to identify a group of teachers who may reasonably be assumed to require targeted professional development. These are the teachers with the lowest estimates of relative effectiveness. The final determination, as well as the specific kind of support needed, requires direct observation of classroom performance and consultation with both the teacher and school administrators. In other words, the use of VAMs does not obviate the need to collect other types of information for the evaluation process.

Most importantly, VAM results should *not* be used as the sole or principal basis for making consequential decisions about teachers (concerning salaries, promotions and sanctions, for example). There are too many pitfalls in making "effective teacher" determinations using the kind of data typically available from school districts. One can imagine, however, an important role for a quantitative component in a thorough teacher evaluation process. Such a process has yet to be implemented. Although improved teacher accountability is a legitimate goal, it is only one of many levers available to states in their quest to enhance the quality of teaching over the long term. A comprehensive and sustained strategy is more likely to be successful than a more narrowly focused initiative. ■

Introduction

The most recent reauthorization of the Elementary and Secondary Education Act, the No Child Left Behind Act (NCLB), has been much more successful than its 1994 predecessor in galvanizing states into action. Undoubtedly, the main reason is the loss in federal aid that states would incur should they fail to comply with NCLB mandates — principally, those relating to schools and teachers. School accountability has a strong empirical component: primarily, a test-score-based criterion of continuous improvement, termed "adequate yearly progress" (AYP).

NCLB also requires states to ensure that there are highly qualified teachers in every classroom, with "highly qualified" defined in terms of traditional criteria such as academic training and fully meeting the state's licensure requirements. Focusing attention on teacher quality has been widely welcomed.[1] Interestingly, in this respect, some states have taken the lead by seeking an empirical basis for evaluating teachers, one that draws on evidence of their students' academic growth.[2] Indeed, so the argument goes, if good teaching is critical to student learning, then can't student learning (or its absence) tell us something about the quality of the teaching they have received? Although the logic seems unassailable, it is far from straightforward to devise a practical system that embodies this reasoning.

Over the past decade or so, a number of attempts to establish a quantitative basis for teacher evaluation have been proposed and implemented. They are usually referred to by the generic term "value-added models," abbreviated "VAMs." Essentially, VAMs combine statistically adjusted test score gains achieved by a teacher's students. Teachers are then compared to other teachers in the district based on these adjusted aggregate gains. Various VAMs differ in the number of years of data they employ, the kinds of adjustments they make, how they handle missing data, and so on.

There is a marked contrast between the enthusiasm of those who accept the claims made about VAMs and would like to use VAMs, on the one hand, and, on the other, the reservations expressed by those who have studied their technical merits. This disjuncture is cause for concern. Because VAMs rely on complex statistical procedures, it is likely that policymakers, education officials, teachers and other stakeholders could all benefit from an understandable guide to the issues raised by the use of VAMs for teacher evaluation. (Although there is also considerable interest in using VAMs for school accountability, we will not address that topic here.[3])

This report is designed to serve as such a guide, reviewing the strengths and weaknesses of VAMs without getting bogged down in methodological matters. It is organized in a Q&A format and draws on recent technical publications, as well as the general statistical literature.[4] The intent is to assist interested parties in their deliberations about improving teacher evaluation and to promote the responsible use of a powerful statistical tool. ■

[1] See for example K. M. Landgraf, *The Importance of Highly Qualified Teachers in Raising Academic Achievement* (Testimony before the Committee on Education and the Workforce, U. S. House of Representatives, April 21, 2004.)

[2] Such evaluations may be used to identify teachers in need of professional development, for administrative purposes (e.g., rewards and sanctions), or both.

[3] There are both similarities and differences in the use of VAMs for school and teacher accountability.

[4] This report draws heavily from D. F. McCaffrey et al., *Evaluating Value-Added Models for Teacher Accountability,* Santa Monica, CA: RAND Corporation, 2003. The most relevant parts of the statistical literature deal with drawing causal inferences from different kinds of studies. The classic reference is W. R. Shadish, T. Cook, and D. T. Campbell, *Experimental and Quasi-Experimental Designs for Generalized Causal Inference,* Boston, MA: Houghton Mifflin Company, 2002.

Questions About Measuring Value-Added

1. Why Is There Such Interest in Value-Added Modeling?

In almost all school districts, teacher evaluation is a notoriously subjective exercise that is rarely directly linked to student achievement. Developers of VAMs argue that their analysis of the changes in student test scores from one year to the next enables them to isolate objectively the contributions of teachers and schools to student learning. If their claims are correct, then we have at hand a wonderful tool for both teacher professional development and teacher evaluation.

One attraction of VAMs is that this approach to accountability differs in a critical way from the adequate yearly progress (AYP) provisions of the NCLB Act. To evaluate AYP, a school must compute for all students in a grade, as well as for various subgroups, the proportions meeting a fixed standard, and then compare these proportions with those obtained in the previous year. A number of observers have pointed out the problems arising from making AYP judgments about schools or teachers on the basis of an absolute standard.[5] The issue, simply, is that students entering with a higher level of achievement will have less difficulty meeting the proficiency standard than those who enter with a lower level. (Specifically, the former may have already met the standard or may be very close to it, so they need to make little or no progress to contribute to the school's target.)

Moreover, AYP comparisons are confounded with differences between the cohorts in successive years — differences that may have nothing to do with the schools being evaluated. For example, this year's entering fourth-graders may be more poorly prepared than last year's fourth-graders, making it more challenging for the school to meet its AYP target.

An alternative view, while recognizing the importance of setting a single goal for all students, holds that meaningful and defensible judgments about teachers or schools should be informed by their contributions to the growth in student achievement and not based solely on the proportions of students who have reached a particular standard. In other words, only by following individual students over time can we really learn anything about the roles of schools and teachers.[6] This seems common-sensical — and VAM appears to make this feasible.

For this reason, many individuals and organizations have seized on VAMs as the "next new thing." There have been many reports, as well as articles in the popular press, that tout VAMs as the best, if not the only, way to carry out fair teacher evaluations.[7]

Such widespread interest in VAMs, not to mention their adoption in a number of districts and states, has spurred a number of technical reviews.[8] These reviews paint a somewhat different

[5] R. L. Linn, "Assessments and Accountability," *Educational Researcher, 29* (2), 4-14, 2000. For a perspective on the experience in England, see L. Olson, "Value Lessons," *Education Week, 23*, 36-40, May 5, 2004.

[6] M. S. McCall, G. G. Kingsbury, and A. Olson, *Individual Growth and School Success,* Lake Oswego, OR: Northwest Evaluation Association, 2004; R. L. Linn, *Rethinking the No Child Left Behind Accountability System* (Paper presented at the Center for Education Policy, No Child Left Behind Forum, Washington, DC, 2004); H. C. Doran and L. T. Izumi, *Putting Education to the Test: A Value-Added Model for California,* San Francisco, CA: Pacific Research Institute, 2004; and D. R. Rogosa, "Myths and Methods: Myths About Longitudinal Research, Plus Supplemental Questions," in J. M. Gottman (Ed.), *The Analysis of Change* (pp. 3-66), Hillsdale, NJ: Lawrence Erlbaum Associates, 1995.

[7] D. Fallon, *Case Study of a Paradigm Shift (The Value of Focusing on Instruction),* Education Commission of the States, Fall Steering Committee Meeting, Nov. 12, 2003; K. Carey, "The Real Value of Teachers: Using New Information About Teacher Effectiveness to Close the Achievement Gap," *Thinking K-16*, *8*, pp. 3-42, Education Trust, Winter 2004; A. B. Bianchi, "A New Look at Accountability: 'Value-Added' Assessment," *Forecast*, 1(*1*), June 2003; M. Raffaele, *Schools See 'Value-Added' Test Analysis as Beneficial,* Retrieved March 19, 2004, from the online edition of the *Pittsburgh Post-Gazette*, 2004; K. Haycock, "The Real Value of Teachers: If Good Teachers Matter, Why Don't We Act Like It?" *Thinking K-16*, 8(*1*), pp. 1-2, Education Trust, Winter 2004; and D. M. Herszenhorn, "Test Scores to Be Used to Analyze Schools' Roles," *New York Times,* June 7, 2005, p. B3.

[8] R. Bock, R. Wolfe, and T. Fisher, *A Review and Analysis of the Tennessee Value-Added Assessment System* (Technical Report), Nashville, TN: Tennessee Office of Education Accountability, 1996; R. Meyer, "Value-Added Indicators of School Performance: A Primer," *Economics of Education Review,* 16, 183-301, 1997; H. Kupermintz, "Teacher Effects and Teacher Effectiveness: A Validity Investigation of the Tennessee Value Added Assessment System," *Educational Evaluation and Policy Analysis, 25*, 287-298, 2003; and McCaffrey et al., 2003.

picture. While acknowledging that VAMs are an important advance in applied measurement, the reviewers all strongly caution against their uncritical application, especially if the results are to have serious consequences for individuals or schools. Ultimately, the concerns are related to questions of proper test use.[9]

2. What Is the Fundamental Concern About VAMs?

The application of most VAMs involves both intricate statistical methodology and knotty questions of interpretation. But before confronting some of the technical issues arising in the use of VAMs in teacher evaluation, it is important to raise a fundamental problem that bedevils any attempt to measure teacher effectiveness.

At the conclusion of a value-added analysis, a number is associated with each teacher. That number, expressed in scale score points, may take on both positive and negative values. It describes how different that teacher's performance is from the performance of the typical teacher, with respect to the average growth realized by the students in their classes. It is often referred to as a measure of "teacher effectiveness." A problem arises because the word "effectiveness" denotes a causal interpretation. That is, the reader is invited to treat those numbers as if, in fact, they unambiguously capture the relative contributions of different teachers to student learning. Thus, if a teacher with an effectiveness of +6 were replaced by a teacher with an effectiveness of only +2, we should expect that the test scores in a typical class would be lower by an average of four points, other things being equal.

Obviously, such a change can never be directly observed because the same class cannot be simultaneously taught full time by two different teachers. So we must somehow infer, from the data we do have, what the relative contributions of different teachers would be. To make the causal interpretation explicit we have to specify the populations under study, describe the nature of the measure(s) employed, and define effectiveness in precise, quantitative terms.[10]

For example, we might want to evaluate all fourth-grade teachers in a particular district, using as our measure the increase in scores on a particular test over the course of the school year. We could define the effectiveness of a teacher as the difference between the average gain that would be obtained by a class taught by this teacher and the average gain that would be obtained by that same class if taught by the average teacher in the district. This would constitute a comparative or relative approach to teacher evaluation.

According to statistical theory, the ideal setting for obtaining proper estimates of teacher effectiveness (as defined above) is a school system in which, for each grade, students are randomly grouped into classes, and teachers in that grade are randomly allocated to those classes. Roughly speaking, randomization levels the playing field for all teachers in that each teacher has an equal chance of being assigned to any class.[11] The data generated in such a setting would allow us to obtain a reasonable estimate of each teacher's effectiveness, as well as a measure of the precision to be attached to the estimate. A finding that the average student growth associated with a particular teacher is significantly greater than the district average would be credible evidence for that teacher's relative effectiveness.

Unfortunately, school systems do not operate by randomization. Many parents have strong opinions about which districts (and even which schools within districts) they want their children to attend, and make corresponding decisions about housing. Within a school, parents often exert influence on the class or teacher to which their child is assigned. Similarly, teachers can sometimes select which district to work in and, by dint of seniority, have some choice in the classes they teach, or even the schools in which they are placed.

[9] American Psychological Association, American Educational Research Association, and National Council on Measurement in Education, *Standards for Educational and Psychological Testing,* Washington, DC: American Psychological Association, 1999.

[10] These issues are explored in greater detail in McCaffrey et al., 2003, pp. 7-15.

[11] Although randomization is an essential component of a proper experiment, there are additional complications in the teacher evaluation setting. See the answer to question 3 for more details.

Since randomization is typically infeasible for the purpose of estimating teacher effects, we must resort to collecting data on teachers and students as they are found in their schools and classrooms. We then use statistical models and procedures to adjust, to the extent possible, for circumstances such as those just described.[12] It is impossible, however, to document and model all such irregular circumstances; yet they may well influence, directly or indirectly, the answers we seek nearly as much as what the teacher actually does in the classroom.

The fundamental concern is that, if making causal attributions is the goal, then no statistical model, however complex, and no method of analysis, however sophisticated, can fully compensate for the lack of randomization.

The problem is that, in the absence of randomization, it is hard to discount alternative explanations for the results that are found. (This explains why many consider randomized experiments the gold standard in scientific work.[13]) Specifically, teacher effects based on statistical estimates may actually represent the combined contributions of many factors in addition to the real teacher contribution we are after. Thus the estimate could be fundamentally off target.[14] Further, it is usually difficult to determine how off target an estimate is. Clearly, substantial discrepancies would seriously undermine the utility of inferences made on the basis of the analysis.

A number of authors have highlighted the distinction between "effects," which are the output of a statistical algorithm, and "effectiveness," which is an interpretation relating to the direct contribution of a teacher to student academic growth.[15] Careful consideration of this distinction in the context of schools brings to the fore the many pitfalls in interpreting "effects" as "effectiveness." This is exactly where the lack of randomization causes difficulties.[16]

Developers of VAM software and those who employ the results rarely acknowledge the implications of the fundamental problem. The assumptions required to justify endowing the estimated teacher effects with a causal interpretation (i.e., treating them as statistically unbiased estimates of teacher effectiveness) are usually not made explicit. Simply said, VAM proponents are behaving as if any statistical bias is too small to worry about. Unfortunately, most of the assumptions made are not directly testable. Thus, the credibility of the causal interpretations, as well as the inferences and actions that follow, must depend on the plausibility of those assumptions. In the context of real-world schools, judging plausibility is a very difficult matter.

3. What Are Some Specific Concerns About Treating Estimated 'Teacher Effects' as Measures of 'Teacher Effectiveness'?

• Inappropriate attribution

Because the ways teachers and students are matched in real schools may be related to the students' potential or rate of growth, teachers can be inappropriately credited or penalized for their students' results. For example, teachers with seniority are usually given more choice in the schools and classes they teach. Suppose they opt to work in schools with better conditions and in classes with students who are better prepared and more engaged. Those students may have a greater intrinsic

[12] In this context, the use of (simple) average gain scores cannot be recommended. More complicated methods are called for.

[13] This point is somewhat controversial. A good general presentation can be found in R. J. Shavelson and L. Towne, *Scientific Research in Education,* Washington, DC: National Academy Press, 2002. For discussion, see M. J. Feuer, L. Towne, and Richard J. Shavelson, "Scientific Culture and Educational Research," *Educational Researcher,* 31(*8*), 4-14, 2002; and J. A. Maxwell, "Causal Explanation, Qualitative Research, and Scientific Inquiry in Education," *Educational Researcher,* 33(*2*), 3-11, 2004.

[14] The technical term for an estimator being off target is that it is "statistically biased." The use of the word "bias" here is different from such everyday meanings as "unfair" or "prejudiced." Rather, it signifies that the differences between the estimator and its target cannot be made to vanish simply by accumulating more data.

[15] For a lucid exposition in the present context, see Kupermintz, 2003.

[16] Random matching of teachers and students would enable us to discount a number of alternative explanations for a finding of wide variation among estimated teacher effects. See Shadish et al., 2002.

rate of growth and, consequently, their teachers' (apparent) effectiveness could be inflated. Conversely, newer and less qualified teachers may be assigned to schools with poorer conditions and to classes with students who are less prepared and less engaged. The (apparent) effectiveness of these teachers will likely be reduced. VAMs generally cannot eliminate these systematic misattributions.

Another, related, issue concerns so-called context effects. Student learning during the year is not just a function of a student's ability and effort, and the teacher's pedagogical skills. It also is affected by such factors as peer-to-peer interactions and overall classroom climate. To be sure, these variables are partially affected by the teacher; but with VAMs, the estimated teacher effect fully incorporates the contributions of all these factors, because there is no other component of the model to capture them. This can also be a source of misattribution.

Further, student learning can be influenced by characteristics of the school, such as the availability of school resources, as well as by both school policies and differential treatment of schools by the district. Because teachers are not randomly distributed across schools, if these factors are not included in the model, then their contributions to student learning are absorbed into the estimated teacher effects.[17]

There is no easy way to address these issues. Including a school model in the VAM system can help somewhat, but may introduce other biases when there is a clustering of teachers of (true) differential effectiveness by school.[18] It is essentially impossible to fully disentangle the contributions of the different factors in order to isolate the teacher's contribution (i.e., obtain a statistically unbiased estimate of a teacher's effectiveness).

• Consequences of missing data

A district database compiled over time will generally have a substantial amount of missing data. Most commonly, the link between a student and a teacher for a given subject and grade is missing. If there are student test data, they can be included in the calculation of the district averages but will not contribute directly to the estimation of teacher effects. If the fact that the link and/or the test score are missing is related to the score the student received, or would have received, then there is some bias in the estimated teacher effects.[19]

• Assumptions underlying the models

Another set of problems arises whenever one relies on mathematical models of real-world phenomena. In one VAM version, for example, it is assumed that a teacher's effect is essentially the same for all of that teacher's students in a given subject and year and, moreover, that this effect persists undiminished into the future for those students. Such assumptions may be more or less plausible, but they do require some validation rather than being accepted uncritically. If these assumptions substantially deviate from reality, the resulting estimates of teacher effects will be biased.[20]

In comparing teachers in a particular grade on the basis of their estimated effects, there is an implicit assumption that they all have been assigned similar academic goals for their classes and have equivalent resources. This flies in the face of the reality that tracking is endemic in schools, particularly in middle schools and above. Students in different classes may be exposed to different material and assigned different end-of-year targets. These differences will influence the estimates of teacher effects. Moreover, different schools in the same district may be employing different curricula or following different reform strategies.

[17] It even appears that students do more poorly in a grade if it is the lowest grade in the school. So, for example, seventh-grade students in a school with only seventh and eighth grades do more poorly on average than students in a school with sixth, seventh, and eighth grades. See W. J. Sanders and S. Horn, "Research Findings from the Tennessee Value-Added Assessment System (TVAAS) Database: Implications for Educational Evaluation and Research," *Journal of Personnel Evaluation in Education, 12*, 247-256, 1998. Again, these contextual factors affect the estimated teacher effects.

[18] McCaffrey et al., 2003.

[19] The technical term is that the missing data are not missing at random. See McCaffrey et al., 2003, for a discussion of the problem.

[20] McCaffrey et al., 2003, develop a model that does not require the assumption that teacher effects persist undiminished into the future. When they apply this model to data, they obtain different results from those obtained with the assumption. However, there is an argument for assuming the simpler model, based on the relationship between the estimated effects and student characteristics. (W. J. Sanders, personal communication, March 18, 2005.)

Another critical decision centers on whether to incorporate student background characteristics in the model. As we shall see, some approaches to VAM do and some don't. Although it is well known that student characteristics are strongly correlated with student attainment, it appears that the correlation is much weaker with changes in attainment. This is an empirical finding, and it may still be the case that bias can be introduced if the model does not capture certain features of the students' demographics.

- Precision of estimates

We have already noted that the randomized experiment is considered to be the gold standard in experimental work. In medicine, for example, randomized clinical trials are almost always required to obtain FDA approval for a new drug or procedure. In such cases, however, the number of alternative treatments to be compared is usually rather small — typically fewer than four. If a large number of patients have been randomly allocated to the different treatments, we can assume that other factors (besides the treatments under investigation) that might affect the study outcomes have been averaged out across the treatments. Consequently, a statistically significant observed difference between a pair of treatments can be reasonably attributed to a real difference in efficacy, because plausible alternative explanations are unconvincing.

Unfortunately, obtaining useful estimates of teacher effects is more problematic — even if random allocation were feasible. The difficulty is that, in the education setting, teachers play the role of treatments. Thus, in a typical district with hundreds of teachers, the amount of information available for each teacher is relatively small, consisting of the data from just a few classes. Some VAMs try to remedy the situation by augmenting the data available for each student by including test scores from previous and future years, as well as from different subjects. While this can help, it does raise other concerns, as we shall see.

With a relatively small number of students contributing to the estimated effect for a particular teacher, the averaging power of randomization can't work for all teachers in a given year. Suppose, for example, that there are a small number of truly disruptive students in a cohort. While all teachers may have an equal chance of finding one (or more) of those students in their class each year, only a few actually will — with potentially deleterious impact on the academic growth of the class in that year. The bottom line is that even if teachers and students come together in more or less random ways, estimated teacher effects can be quite variable from year to year.

In summary, given sufficient data, a reasonable statistical model, and enough computing power, it is always possible to produce estimates of what the model designates as teacher effects. These estimates, however, capture the contributions of a number of factors, those due to teachers being only one of them. Moreover, the estimates may be quite volatile. So treating estimated teacher effects as accurate indicators of teacher effectiveness is problematic. Much more needs to be known about these kinds of data and the properties of the models in different, commonly occurring situations before there can be agreement on whether it is generally possible to isolate teachers' contributions to student learning, and have the confidence to carry out actions on that basis.

4. What Value-Added Models Are Now in Use?

There are several VAMs in circulation. They are similar in that they are purely statistical in nature and rely solely on student test scores, and not on other measures of student learning or such sources of information as interviews with students, teachers or administrators. Users of any of these models must confront the fundamental problem that the lack of random pairings among students and teachers makes causal attributions very problematic. The models do differ, however, in their structure and the kinds of assumptions they make.[21]

[21] McCaffrey et al., 2003.

The outcome of applying any of these models is that some number of teachers are identified as being significantly better or worse than average. Not surprisingly, findings can differ across approaches. Some VAMs are listed below:

- EVAAS (the Educational Value-Added Assessment System) is the best known and most widely used VAM. It was developed by William Sanders and his associates for use in Tennessee and has been in place there since 1993. Since then, it has been considered and, in some cases, adopted by districts in other states.[22] (A fuller description of the EVAAS can be found in the answer to Question 5.)
- DVAAS (the Dallas Value-Added Accountability System) is a widely cited alternative to the EVAAS and has been employed by the Dallas school system for a number of years.[23] It uses a value-added criterion to identify highly effective teachers, as well as those in need of support. The DVAAS differs from the EVAAS in four important ways. First, it does use student-level characteristics to adjust student test scores prior to analysis. Second, it only models the relationship between adjusted test scores in adjacent grades (as opposed to combining data across several grades). Third, it doesn't directly model gains in the adjusted test scores but, rather, a more general structural connection between them.[24] Finally, the model includes not only a teacher's contribution to student achievement but also a number of other factors that are intended to account for the influence of the school on student achievement.
- Another alternative, REACH (Rate of Expected Academic Change), has been suggested by Doran and Izumi for use in California. Their test-based criterion measures student progress toward meeting a proficiency standard. Thus, each student's growth is measured against a goal rather than against the growth of other students. Doran and Izumi argue that this is a more constructive way of measuring AYP. They note that such a value-added criterion could also be used to evaluate teacher effectiveness, but do not suggest a particular model for obtaining estimates of teacher effects.[25]

Other VAMs have been proposed but used only in a research context.[26] Since EVAAS is the most widely used model for evaluating teacher effectiveness, this paper will henceforth focus on that model.

5. How Does EVAAS Work?

The building blocks of the EVAAS model are rather simple, with the complexity arising in the aggregation of data across students, subjects and years. We confine ourselves here to a summary that is necessarily incomplete.[27] The basic model is an equation that expresses the score of a student at the end of a particular grade in a particular year as the sum of three components:

- District average for that grade and year
- Class (teacher) effect
- Systematic and unsystematic variations

Thus, the essential difference between the student's score and the average score in the district is attributed to a "class effect" plus the combined

[22] The Ohio Partnership for Accountability, including all 51 of Ohio's schools of education, the State Department of Education, and the Board of Regents, has announced a project to use value-added teacher effectiveness data to better understand, study and improve university teacher preparation programs. Ohio is using a variant of the EVAAS. With cooperation of the state's teachers unions, Ohio's project is the first statewide effort of its kind. Over the next five years, Ohio researchers will study the math and reading scores of the students of both new and veteran teachers as a means to evaluate the quality of teacher preparation and to identify the most effective practices and policies. The Milken Foundation's Teacher Advancement Program, operating in schools in Arizona, Colorado, Indiana, Louisiana and South Carolina, includes the value added by teachers to their students' achievement in its school reform model. Other districts using value-added measures of teacher effectiveness to improve teaching and learning include the Minneapolis Public Schools, Guilford County, North Carolina, as well as a number of districts in Pennsylvania.

[23] W. Webster and R. Mendro, "The Dallas Value-Added Accountability System," in J. Millman (Ed.), Grading Teachers, Grading Schools: Is Student Achievement a Valid Evaluation Measure? (pp. 81-99). Thousand Oaks, CA: Corwin Press, Inc, 1997.

[24] The technical term for this type of model is "analysis of covariance."

[25] Doran and Izumi, 2004.

[26] B. Rowan, R. Correnti, and R. J. Miller, "What Large-Scale Survey Research Tells Us About Teacher Effects on Student Achievement: Insights from the Prospects Study of Elementary Schools," Teachers College Record, 104, 1525-1567, 2002.

[27] The best technical description of EVAAS can be found in W. L. Sanders, A. Saxton, and B. Horn, "The Tennessee Value-Added Assessment System: A Quantitative Outcomes-Based Approach to Educational Assessment," in J. Millman (Ed.), Grading Teachers, Grading Schools: Is Student Achievement a Valid Evaluation Measure? (pp. 137-162), Thousand Oaks, CA: Corwin Press, Inc, 1997.

contributions of unspecified variations, including measurement errors. It is assumed that the class effect is the same for all the students in the class and attributable to the teacher of the class.[28] (Henceforth, we refer to it as a teacher effect.) When the student moves to the next year and the next grade, the model then has four components:

- District average for that grade and year
- Teacher effect for that year
- Teacher effect from the previous year
- Systematic and unsystematic variations

Note that it is assumed that the teacher effect for the previous year persists undiminished into the current year and that the components of the unspecified variations in the two years are unrelated to each other. Finally, if we subtract the first-year score from the second-year score, we obtain the gain made by the student. According to the model, this must be the sum of:

- Average gain for that grade in the district
- Teacher effect of the second-year teacher
- The two error terms

That is, ignoring the error terms, the teacher effect in the second year is the difference between the gain experienced by the student in that year and the average gain in the district for that same year. This formulation is intuitively plausible and attractive.

It is possible to add equations for the data from subsequent years. Sanders uses the term "layered model" to capture the notion that the data from each succeeding year are added to those from the previous years. In a typical application, students may contribute as many as five years of data. Moreover, student gains in different subjects are included in the EVAAS model, with each subject and year assigned its own equation. It is not hard to see why the database to support the analysis is both large and complex, as it must maintain multiple links between students and teachers over different subjects and years.

The estimate of a teacher effect is based on many different elements, including the growth in learning (as measured by an increase in test scores) of the students in the teacher's classes over a number of years, adjusted for the effects of previous teachers of those students; the growth of the teacher's students in subsequent years; and the achievements of those students in different subjects over a number of years, all appropriately adjusted for the contributions of those students' other teachers.[29] It is virtually impossible to visualize how all these elements are combined to yield an estimate of the teacher's value.

Sanders argues that there is no need to include student characteristics (e.g., gender, race, socioeconomic status, and so on) in the model. His rationale is that, while there are substantial correlations between these characteristics and the current level of achievement, the correlations of these characteristics with gains are essentially zero. However, this is an assertion based on his reading of the data and not a mathematical certainty. This issue has been subjected to empirical examination and has not been found to be universally valid.[30] For this reason, some argue that fairer estimates of teacher effects will result if student characteristics are included in the model. Recently, it has been shown how this can be done as part of the EVAAS approach.[31] Unfortunately, this is not the end of the story, since the issues raised so far are still relevant to the proper interpretation of the resulting estimates.

The EVAAS model is very efficient in that it makes use of all the test information available for a given cohort of students within a moving five-year window. The estimation algorithms are able to handle various patterns of missing data so that

[28] The identification of the class effect with teacher effectiveness conflates two separate steps: First, endowing a statistical quantity (class effect) with a causal interpretation and, second, attributing the causal contribution of the class entirely to the teacher. See H. I. Braun, "Value-Added Modeling: What Does Due Diligence Require?" in R. Lissitz (Ed.), Value Added Models in Education: Theory and Applications, pp. 19-39. Maple Grove, MN: JAM Press, 2005.

[29] Although there is a separate equation for each subject and year, all the equations for a given cohort are tied together through another model feature (covariance matrices) that captures the fact that test scores for a given student over time and across subjects are statistically related to one another. This knitting together of disparate test scores distinguishes EVAAS from approaches based on simple comparisons of average gains across classes.

[30] McCaffrey et al., 2003.

[31] D. Ballou, W. Sanders, and P. Wright, "Controlling for Students' Background in Value-Added Assessment for Teachers," *Journal of Educational and Behavioral Statistics*, 2004.

if data on a particular student are unavailable in a given year, the remaining data can be incorporated into the analysis. In particular, student data that are not linked to a specific teacher still contribute to the estimation of the district averages. Sanders is correct in citing this as an advantage of his approach. However, as indicated earlier, if the patterns of missing data are related to student performance or teacher effectiveness, then systematic errors can be introduced into the estimated teacher effects. Sanders claims that by incorporating information over time and across subjects, the estimates generated by the EVAAS model are relatively unaffected by unusual patterns of missing data. Again, this claim requires empirical validation.

The principal output of an EVAAS analysis is a set of estimated teacher effects.[32] These estimates have well-established statistical properties. From the various studies they conducted, Sanders and his associates observed some heterogeneity among the estimated effects, which they interpreted as indicating real differences in teacher effectiveness.[33] (Indeed, they argue, as do many others, that teachers are the main source of variation in student gains.) Empirically, however, no more than a third of the teachers in a district have been reliably shown by EVAAS to be different from the average. Often the fraction is much smaller.

6. What Are Some of the Issues in Using Student Achievement Data in Teacher Evaluation?

It seems quite reasonable to judge teachers on the basis of their contributions to student learning.[34] Operationally, this means relying on scores obtained from standardized tests. One of the attractive properties of these scores is that they are hard numbers, as opposed to other qualities of students that we might be interested in documenting, such as engagement and enthusiasm, which are more difficult to measure.

We should recognize, however, that test scores are the final result of a complex process that involves translating state standards into test specifications and those, in turn, into test items assembled in a particular way to constitute an operational assessment. At each stage, design decisions are made on the basis of professional judgment, balancing substantive and psychometric considerations against constraints of cost, testing time, and so on. Good practice requires that such test characteristics as the nature of the scale score and the validity of the test be examined in light of the proposed uses of the test scores.

The Score Scale. If different forms of a test are used for a particular grade each year, as is usually recommended, then the scores in the same grade from different years must be put on the same scale so that gains in different years are comparable. This involves a statistical procedure called (horizontal) equating that is common practice. Though it is usually done well, it does introduce uncertainty into the reported scores.[35]

Of greater concern is that, in some applications of VAMs, student scores over as many as five grades may be included in the database. These scores are not obtained from a single test form administered in all the grades but from a number of test forms that, presumably, have each been designed to be grade-appropriate. Consequently, as we move to higher grades, the detailed specifications that govern the construction of each test will reflect the greater dimensionality and expanded knowledge base of the subject. This evolution in complexity is masked somewhat because test results are summarized in a single total score, and secondary analyses for evaluation typically utilize this total score.[36]

[32] For a given year and subject, a teacher can be associated with as many as three estimated effects, one for each of three successive cohorts. In Tennessee, schools are provided with the average of these three effects. In the following year, the data window shifts: the earliest cohort is dropped, effects for the two remaining cohorts are reestimated, an effect for a new cohort is obtained, and a new, three-cohort average is calculated.

[33] W. L. Sanders and J. C. Rivers, *Cumulative and Residual Effects of Teachers on Future Student Academic Achievement,* Knoxville, TN: University of Tennessee Value-Added Research Center, 1996; and Sanders and Horn, 1998. Estimated teacher effects are normatively defined; i.e., as deviations from the average teacher in the district. As such, teacher effects cannot be compared across districts that were separately analyzed.

[34] Educational Testing Service, *Where We Stand on Teacher Quality* (Teacher Quality Series), Princeton, NJ: Educational Testing Service, 2004.

[35] M. J. Kolen and R. L. Brennan, *Test Equating: Methods and Practices,* New York, NY: Springer, 1995.

[36] The total score is, typically, not a simple sum of the number of correct responses. It is, rather, a weighted composite constructed from subscale scores derived in turn from complex measurement models applied to the raw test data.

The total scores on the different instruments are usually placed on a common scale through another statistical procedure called (vertical) scaling, which introduces additional uncertainty into the process.[37] Aside from the technical aspects of vertical scaling, there is a question of what it means to put, say, third-grade and seventh-grade mathematics scores on the same scale. In particular, should we treat a 20-point gain at the low end of the scale as equivalent to a 20-point gain at the upper end? Doing so requires making very strong assumptions about the nature of growth over the grade span of interest. More to the point, should we expect that the average teacher teaching a typical class would obtain the same (relative) growth irrespective of the grade? This is unlikely given the increasing complexity of the construct at higher grades. Although the question could be addressed empirically, this appears not to have been done.

Validity. Whether test scores actually measure what they are intended to measure is the basic concern of validity.[38] Typically, state content standards are broad, ambitious and often ambiguous. The degree of articulation between tests and the standards varies among states and even across subjects and grades within a state. Indeed, reviews of state tests often find that they don't measure some of the content standards at all and some only superficially, focusing instead on those aspects of the standards that can be probed with multiple-choice questions.[39] For example, a standard in language arts addressing the ability of a student to write a well-crafted essay should be measured by having the student write an essay. Most would agree that a multiple-choice test falls short. These considerations give rise to two related concerns: First, that "teaching to the test" may result in increased test scores that do not generalize to gains in the broader achievement domain that the test is intended to measure. Second, that teachers who do try to teach the full curriculum may find their students not gaining as much as others, whose teachers resort to some form of teaching to the test.[40] That is, the test may not be sensitive to the full range of students' learning gains. It is possible that this problem is exacerbated by the use of test scores obtained through a vertical scaling procedure.[41]

In sum, a rigorous evaluation of the validity of the assessment battery used by a state is an essential foundation for appropriate test use. The alignment of the test with the corresponding standards, as well as the shift in the meaning of the score scale across grades, should be taken into account in deciding how to best use test scores. Given the current state of the art, caution is warranted. Policymakers should have technical support in deciding whether the test score scale can support the interpretive burden placed on it — and moderate their use of VAM results accordingly.

To cite one example, suppose that, on average, reported score gains are typically smaller the higher the students' initial scores. In that case, two teachers of equal effectiveness, but assigned over time to classes with substantially different distributions of initial scores, can find themselves with quite different estimated effects. If these are interpreted as indicators of differential effectiveness, then teachers are ill-served by the process. Of course, these difficulties would be mitigated somewhat if we only compared teachers at the same grade level.[42]

[37] D. J. Harris et al., *Vertical Scales and the Measurement of Growth*, paper presented at the Annual Meeting of the National Council on Measurement in Education, San Diego, CA, 2004. McCaffrey et al., 2003, also discuss some of the measurement issues. For a more accessible discussion, see D. Ballou, "Sizing Up Test Scores, *Education Next, 2002*. Retrieved May 26, 2004, from http://www.educationnext.org/20022/10.html.

[38] See for example L. J. Cronbach, "Five Perspectives on Validity Argument," In H. Wainer and H. I. Braun (Eds.), *Test Validity* (pp. 3-17). Hillsdale NJ: Lawrence Erlbaum Associates, Inc., 1988. The classic reference is S. Messick, "Validity," In R. L. Linn (Ed.), *Educational Measurement* (3rd Ed., pp. 13-103). New York, NY: Macmillan Publishing Co., 1989.

[39] American Federation of Teachers, "Executive Summary," *Making Standards Matter 2001*. Retrieved May 25, 2004, from http://www.aft.org/edissues/standards/MSM2001/downloads/execsummary.pdf. For an interesting perspective, see G. W. Bracey, *A Review of: The State of State Standards,* (Thomas B. Fordham Foundation, January 2000), (No. CERAI-00-07), Milwaukee, WI: Center for Education Research, Analysis, and Innovation, Feb. 2, 2000.

[40] For two of many views, see L. Bond, "Teaching to the Test," *Carnegie Perspectives*, 2004. Retrieved Aug. 3, 2004, from http://www.carnegiefoundation.org/perspectives/perspectives2004.Apr.htm, and W. J. Popham, "Teaching to the Test?" *Educational Leadership, 58*(6), 16-20, 2001.

[41] W. H. Schmidt, R. Houang, and C. C. McKnight, "Value-Added Research: Right Idea but Wrong Solution?," In R. Lissitz (Ed.), *Value-added Models in Education: Theory and Applications* (pp. 145-164). Maple Grove, MN: JAM Press, 2005.

[42] Some may think (wrongly, as it happens) that such difficulties are particularly severe with external tests. Actually, local tests can be deficient with respect to both reliability and validity. They are even more problematic with respect to comparability across schools or districts. Thus, there is probably no alternative but to use externally developed tests for teacher evaluation.

7. Where Do We Stand?

There is progress and promise in that:

- VAM moves the discussion about teacher quality to where it belongs: centered on increasing student learning as the primary goal of teaching. It can also enhance the teacher evaluation process by introducing a quantitative component, as well as by forcing us to reexamine questions of fairness and proper test use. These are major steps in the right direction.
- By utilizing measures related to individual student growth, VAM provides a more defensible foundation for teacher evaluation than is offered by methods based on the proportion of students meeting a fixed standard of performance.
- There have already been a number of investigations of different VAMs in a variety of settings. They have begun to give us a clearer picture of the strengths and limitations of the various approaches.

There are appropriate uses of VAM results, such as:

- Identifying teachers who are most likely to require professional development and who should be interviewed and/or observed to determine the particular kinds of support that would be most helpful. This screening strategy would help in allocating scarce resources to those teachers in greatest need.[43]
- Identifying schools that may be underperforming and should be audited to determine whether they are in need of specific kinds of assistance.[44]

There are cautions, such as:

- VAM results should not serve as the sole or principal basis for making consequential decisions about teachers. There are many pitfalls to making causal attributions of teacher effectiveness on the basis of the kinds of data available from typical school districts. We still lack sufficient understanding of how seriously the different technical problems threaten the validity of such interpretations.
- Although we can all agree on the importance of teacher evaluation, identifying precisely which teachers are deserving of commendation and which are in need of focused professional support is another matter entirely. Unfortunately, extreme ranks, those near the top or near the bottom, are very unreliable.[45]
- The use of VAMs should not block the examination of the appropriateness or desirability of including other measures, in addition to student test scores, in teacher evaluation.[46] Moreover, we must recognize that statistical models cannot identify the strategies and practices teachers employ. Expert observation, portfolio reviews, conversations with teachers, and so forth, are essential to making informed judgments about whether one teacher truly excels or whether another really needs support. School leaders should also become more skilled in recognizing the kinds of assistance needed by individual teachers. ■

[43] This type of use was cited in Sanders and Horn, 1998, as the primary function of EVAAS and has been carried out in some districts in Tennessee. Apparently, Dallas has also made good use of its VAM results in building system capacity through targeted professional development. In this regard, see Webster and Mendro, 1997.

[44] Underperforming schools could be identified by looking for clusters of teachers with low estimated effectiveness or by carrying out school-level value-added analysis.

[45] J. Lockwood, T. Louis, and D. F. McCaffrey, “Uncertainty in Rank Estimation: Implications for Value-Added Modeling Accountability Systems,” *Journal of Educational and Behavioral Statistics*, *27*, 255-270, 2002.

[46] In general, the use of multiple sources of information is preferable to the use of a single measure. At the same time, it is important to note that test scores have been subject to much greater scrutiny, and their properties more thoroughly documented, than have other possible measures such as those obtained through direct observation.

Epilog

As a nation, we have come to the realization that good teaching really does matter. Building a credible statistical basis for teacher evaluation would be an important advance that could contribute, in the long run, to improved teaching and learning. But the evaluation process should be sufficiently rich to do justice to the complex, multifaceted activity that constitutes teaching. Reliance on a single, statistical measure cannot be recommended. That said, there is certainly an important role for VAMs to play. Indeed, the logic behind using VAMs is compelling, and VAM-based approaches to teacher accountability have gained numerous adherents in many states. In view of the methodological issues that have surfaced, however, it is critical that further investigations of various approaches be carried out. Fortunately, there is substantial activity in this area.[47]

It may well be that we can never rigorously justify treating estimated teacher effects as accurate indicators of teacher contributions to student learning. Nonetheless, districts employing VAM results in sensible ways might, over time, experience greater improvements in student scores than other comparable districts not using VAMs. That possibility can be investigated using a randomized experiment conducted at the district level. Indeed, it has been argued that such a study would yield results that are more directly related to policymaker concerns than are attempts to validate the causal interpretation of VAM output.[48]

Policymakers should not ignore the technical aspects of VAMs. The concerns that have been raised are central to the proposed use of VAM results in teacher evaluation. An early objection to the EVAAS system was that it was too difficult to understand and thus shouldn't be used to make decisions about teachers. The response was that one didn't have to understand how a car works in order to drive it. That argument seemed to carry the day. However, in view of more recent critiques, which are only summarized here, perhaps the metaphor should be reexamined.

Certainly, one needn't understand how a car works while driving it under the conditions it was designed for. But if there are plans to drive it under nonstandard conditions, say on a beach, it is only prudent to inquire first about the capabilities of the drive train and the transmission before setting off. Similarly, the statistical models underlying EVAAS were originally developed for use in settings, such as agriculture, in which randomized experiments and sufficient data are the norm. Thus, endowing statistical estimates with causal interpretations is relatively straightforward. But taking that same methodology off-road, so to speak, in circumstances with multiple sources of selection bias (and perhaps less data than desired), demands a careful look under the hood. This is just due diligence.

Finally, raising the quality of teaching will require more than instituting better accountability. At the least, jurisdictions implementing VAMs should also be building capacity to help those teachers who are identified as needing improvement. But greater effort is called for: States and districts have many levers at their disposal, among them the improvement of teacher training, standards for licensure, effective mentoring for new teachers, policies regarding the assignment of teachers to schools and to classrooms within schools, more equitable distribution of resources, targeted professional development, as well as higher salaries and differentiated pay schedules. A coherent, sustained and systemic initiative that involves many, if not all, of these levers will surely meet with greater success than a narrow effort focused on just one. ■

[47] See, for example, H. Wainer, "Introduction to the Value-Added Assessment," *Journal of Educational and Behavioral Statistics* (Special Issue), Vol. 29, 1-3, 2004; and R. Lissitz (Ed.), *Value-Added Models in Education: Theory and Applications,* Maple Grove, MN: JAM Press, 2005.

[48] See D. B. Rubin, E. A. Stuart, and E. L.Zanutto, "A Potential Outcomes View of Value-Added Assessment in Education," *Journal of Educational and Behavioral Statistics,* Vol. 29, 103-116, 2004.

ETS®

I.N. 730194

0700