

, 05/09/2011

Leading mathematician debunks 'value-added'

By Valerie Strauss

--This was written by John Ewing, president of Math for America, a nonprofit organization dedicated to improving mathematics education in U.S. public high schools by recruiting, training and retaining great teachers. This article originally appeared in the May Notices of the American Mathematics Society. It gives a comprehensive look at the history, current use and problems with the value-added model of assessing teachers. It is long but well worth your time.

By John Ewing

Mathematicians occasionally worry about the misuse of their subject. G. H. Hardy famously wrote about mathematics used for war in his autobiography, *A Mathematician's Apology* (and solidified his reputation as a foe of applied mathematics in doing so). More recently, groups of mathematicians tried to organize a boycott of the Star Wars [missile defense] project on the grounds that it was an abuse of mathematics. And even more recently some fretted about the role of mathematics in the financial meltdown.

But the most common misuse of mathematics is simpler, more pervasive, and (alas) more insidious: mathematics employed as a rhetorical weapon—an intellectual credential to convince the public that an idea or a process is “objective” and hence better than other competing ideas or processes. This is mathematical intimidation. It is especially persuasive because so many people are awed by mathematics and yet do not understand it—a dangerous combination.

The latest instance of the phenomenon is valued-added modeling (VAM), used to interpret test data. Value-added modeling pops up everywhere today, from newspapers to television to political campaigns. VAM is heavily promoted with unbridled and uncritical enthusiasm by the press, by politicians, and even by (some) educational experts, and it is touted as the modern, “scientific” way to measure educational success in everything from charter schools to individual teachers.

Yet most of those promoting value-added modeling are ill-equipped to judge either its effectiveness or its limitations. Some of those who are equipped make extravagant claims without much detail, reassuring us that someone has checked into our concerns and we shouldn't worry. Value-added modeling is promoted because it has the right pedigree — because it is based on “sophisticated mathematics.” As a consequence, mathematics that ought to be used to illuminate ends up being used to intimidate. When that happens, mathematicians have a responsibility to speak out.

Background

Value-added models are all about tests—standardized tests that have become ubiquitous in K–12 education in the past few decades. These tests have been around for many years, but their scale, scope, and potential utility have changed dramatically.

Fifty years ago, at a few key points in their education, schoolchildren would bring home a piece of paper that showed academic achievement, usually with a percentile score showing where they landed among a large group. Parents could take pride in their child's progress (or fret over its lack); teachers could sort students into those who excelled and those who needed remediation; students could make plans for higher education.

Today, tests have more consequences. “No Child Left Behind” mandated that tests in reading and mathematics be administered in grades 3–8. Often more tests are given in high school, including high-stakes tests for graduation.

With all that accumulating data, it was inevitable that people would want to use tests to evaluate everything educational—not merely teachers, schools, and entire states but also new curricula, teacher training programs, or teacher selection criteria. Are the new standards better than the old? Are experienced teachers better than novice? Do teachers need to know the content they teach?

Using data from tests to answer such questions is part of the current “student achievement” ethos—the belief that the goal of education is to produce high test scores. But it is also part of a broader trend in modern

society to place a higher value on numerical (objective) measurements than verbal (subjective) evidence. But using tests to evaluate teachers, schools, or programs has many problems. (For a readable and comprehensive account, see [Koretz 2008].) Here are four of the most important problems, taken from a much longer list.

1. **Influences.** Test scores are affected by many factors, including the incoming levels of achievement, the influence of previous teachers, the attitudes of peers, and parental support. One cannot immediately separate the influence of a particular teacher or program among all those variables.
2. **Polls.** Like polls, tests are only samples. They cover only a small selection of material from a larger domain. A student's score is meant to represent how much has been learned on all material, but tests (like polls) can be misleading.
3. **Intangibles.** Tests (especially multiple-choice tests) measure the learning of facts and procedures rather than the many other goals of teaching. Attitude, engagement, and the ability to learn further on one's own are difficult to measure with tests. In some cases, these "intangible" goals may be more important than those measured by tests. (The father of modern standardized testing, E. F. Lindquist, wrote eloquently about this [Lindquist 1951]; a synopsis of his comments can be found in [Koretz 2008, 37].)
4. **Inflation.** Test scores can be increased without increasing student learning. This assertion has been convincingly demonstrated, but it is widely ignored by many in the education establishment [Koretz 2008, chap. 10]. In fact, the assertion should not be surprising. Every teacher knows that providing strategies for test-taking can improve student performance and that narrowing the curriculum to conform precisely to the test ("teaching to the test") can have an even greater effect. The evidence shows that these effects can be substantial: One can dramatically increase test scores while at the same time actually decreasing student learning. "Test scores" are not the same as "student achievement."

This last problem plays a larger role as the stakes increase. This is often referred to as Campbell's Law: "*The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to measure*" [Campbell 1976]. In its simplest form, this can mean that high-stakes tests are likely to induce some people (students, teachers, or administrators) to cheat ... and they do [Gabriel 2010].

But the more common consequence of Campbell's Law is a distortion of the education experience, ignoring things that are not tested (for example, student engagement and attitude) and concentrating on precisely those things that are.

Value-Added Models

In the past two decades, a group of statisticians has focused on addressing the first of these four problems. This was natural. Mathematicians routinely create models for complicated systems that are similar to a large collection of students and teachers with many factors affecting individual outcomes over time.

Here's a typical, although simplified, example, called the "split-plot design." You want to test fertilizer on a number of different varieties of some crop. You have many plots, each divided into subplots. After assigning particular varieties to each subplot and randomly assigning levels of fertilizer to each whole plot, you can then sit back and watch how the plants grow as you apply the fertilizer. The task is to determine the effect of the fertilizer on growth, distinguishing it from the effects from the different varieties. Statisticians have developed standard mathematical tools (mixed models) to do this.

Does this situation sound familiar? Varieties, plots, fertilizer ...students, classrooms, teachers?

Dozens of similar situations arise in many areas, from agriculture to MRI analysis, always with the same basic ingredients—a mixture of fixed and random effects—and it is therefore not surprising that statisticians

suggested using mixed models to analyze test data and determine “teacher effects.”

This is often explained to the public by analogy. One cannot accurately measure the quality of a teacher merely by looking at the scores on a single test at the end of a school year. If one teacher starts with all poorly prepared students, while another starts with all excellent, we would be misled by scores from a single test given to each class.

To account for such differences, we might use two tests, comparing scores from the end of one year to the next. The focus is on how much the scores increase rather than the scores themselves. That’s the basic idea behind “value added.” But value-added models (VAMs) are much more than merely comparing successive test scores.

Given many scores (say, grades 3–8) for many students with many teachers at many schools, one creates a mixed model for this complicated situation. The model is supposed to take into account all the factors that might influence test results — past history of the student, socioeconomic status, and so forth. The aim is to predict, based on all these past factors, the growth in test scores for students taught by a particular teacher. The actual change represents this more sophisticated “value added”— good when it’s larger than expected; bad when it’s smaller.

The best-known VAM, devised by William Sanders, is a mixed model (actually, several models), which is based on Henderson’s mixed-model equations, although mixed models originate much earlier [Sanders 1997]. One calculates (a huge computational effort!) the best linear unbiased predictors for the effects of teachers on scores. The precise details are unimportant here, but the process is similar to all mathematical modeling, with underlying assumptions and a number of choices in the model’s construction.

History

When value-added models were first conceived, even their most ardent supporters cautioned about their use [Sanders 1995, abstract]. They were a new tool that allowed us to make sense of mountains of data, using

mathematics in the same way it was used to understand the growth of crops or the effects of a drug. But that tool was based on a statistical model, and inferences about individual teachers might not be valid, either because of faulty assumptions or because of normal (and unexpected) variation.

Such cautions were qualified, however, and one can see the roots of the modern embrace of VAMs in two juxtaposed quotes from William Sanders, the father of the value-added movement, which appeared in an article in *Teacher Magazine* in the year 2000. The article's author reiterates the familiar cautions about VAMs, yet in the next paragraph seems to forget them:

Sanders has always said that scores for individual teachers should not be released publicly. "That would be totally inappropriate," he says. "This is about trying to improve our schools, not embarrassing teachers. If their scores were made available, it would create chaos because most parents would be trying to get their kids into the same classroom."

Still, Sanders says, it's critical that ineffective teachers be identified. "The evidence is overwhelming," he says, "that if any child catches two very weak teachers in a row, unless there is a major intervention, that kid never recovers from it. And that's something that as a society we can't ignore" [Hill 2000].

Over the past decade, such cautions about VAM slowly evaporated, especially in the popular press. A 2004 article in *The School Administrator* complains that there have not been ways to evaluate teachers in the past but excitedly touts value added as a solution:

"Fortunately, significant help is available in the form of a relatively new tool known as value-added assessment. Because value-added isolates the impact of instruction on student learning, it provides detailed information at the classroom level. Its rich diagnostic data can be used to improve teaching and student learning. It can be the basis for a needed improvement in the calculation of adequate yearly progress. In time, once teachers and administrators grow comfortable with its fairness, value-

added also may serve as the foundation for an accountability system at the level of individual educators [Hershberg 2004, 1]."

And newspapers such as *The Los Angeles Times* get their hands on seven years of test scores for students in the L.A. schools and then publish a series of exposés about teachers, based on a value-added analysis of test data, which was performed under contract [Felch 2010]. The article explains its methodology:

"The Times used a statistical approach known as value-added analysis, which rates teachers based on their students' progress on standardized tests from year to year. Each student's performance is compared with his or her own in past years, which largely controls for outside influences often blamed for academic failure: poverty, prior learning and other factors.

Though controversial among teachers and others, the method has been increasingly embraced by education leaders and policymakers across the country, including the Obama administration."

It goes on to draw many conclusions, including:

"Many of the factors commonly assumed to be important to teachers' effectiveness were not. Although teachers are paid more for experience, education and training, none of this had much bearing on whether they improved their students' performance."

The writer adds the now-common dismissal of any concerns:

"No one suggests using value-added analysis as the sole measure of a teacher. Many experts recommend that it count for half or less of a teacher's overall evaluation."

"Nevertheless, value-added analysis offers the closest thing available to an objective assessment of teachers. And it might help in resolving the greater mystery of what makes for effective teaching, and whether such skills can be taught."

The article goes on to do exactly what it says “no one suggests” — it measures teachers solely on the basis of their value-added scores.

What Might Be Wrong with VAM?

As the popular press promoted value-added models with ever-increasing zeal, there was a parallel, much less visible scholarly conversation about the limitations of value-added models. In 2003 a book with the title *Evaluating Value-Added Models for Teacher Accountability* laid out some of the problems and concluded:

“The research base is currently insufficient to support the use of VAM for high-stakes decisions. We have identified numerous possible sources of error in teacher effects and any attempt to use VAM estimates for high-stakes decisions must be informed by an understanding of these potential errors [McCaffrey 2003, xx].”

In the next few years, a number of scholarly papers and reports raising concerns were published, including papers with such titles as *“The Promise and Peril of Using Valued-Added Modeling to Measure Teacher Effectiveness”* [RAND, 2004], *“Re-Examining the Role of Teacher Quality in the Educational Production Function”* [Koedel 2007], and *“Methodological Concerns about the Education Value-Added Assessment System”* [Amrein-Beardsley 2008].

What were the concerns in these papers? Here is a sample that hints at the complexity of issues.

- In the real world of schools, data is frequently missing or corrupt. What if students are missing past test data? What if past data was recorded incorrectly (not rare in schools)? What if students transferred into the school from outside the system?
- The modern classroom is more variable than people imagine. What if students are team-taught? How do you apportion credit or blame among various teachers? Do teachers in one class (say mathematics) affect the learning in another (say science)?

- Every mathematical model in sociology has to make rules, and they sometimes seem arbitrary. For example, what if students move into a class during the year? (Rule: Include them if they are in class for 150 or more days.) What if we only have a couple years of test data, or possibly more than five years? (Rule: The range three to five years is fixed for all models.) What's the rationale for these kinds of rules?

- Class sizes differ in modern schools, and the nature of the model means there will be more variability for small classes. (Think of a class of one student.) Adjusting for this will necessarily drive teacher effects for small classes toward the mean. How does one adjust sensibly?

- While the basic idea underlying value-added models is the same, there are in fact many models. Do different models applied to the same data sets produce the same results? Are value-added models "robust"?

- Since models are applied to longitudinal data sequentially, it is essential to ask whether the results are consistent year to year. Are the computed teacher effects comparable over successive years for individual teachers? Are value-added models "consistent"?

These last two points were raised in a research paper [Lockwood 2007] and a recent policy brief from the Economic Policy Institute, "*Problems with the Use of Student Test Scores to Evaluate Teachers*", which summarizes many of the open questions about VAM:

"For a variety of reasons, analyses of VAM results have led researchers to doubt whether the methodology can accurately identify more and less effective teachers. VAM estimates have proven to be unstable across statistical models, years, and classes that teachers teach. One study found that across five large urban districts, among teachers who were ranked in the top 20% of effectiveness in the first year, fewer than a third were in that top group the next year, and another third moved all the way down to the bottom 40%. Another found that teachers' effectiveness ratings in one year could only predict from 4% to 16% of the variation in such ratings in the following year.

“Thus, a teacher who appears to be very ineffective in one year might have a dramatically different result the following year. The same dramatic fluctuations were found for teachers ranked at the bottom in the first year of analysis. This runs counter to most people’s notions that the true quality of a teacher is likely to change very little over time and raises questions about whether what is measured is largely a “teacher effect” or the effect of a wide variety of other factors [Baker 2010, 1].”

In addition to checking robustness and stability of a mathematical model, one needs to check validity. Are those teachers identified as superior (or inferior) by value-added models actually superior (or inferior)? This is perhaps the shakiest part of VAM. There has been surprisingly little effort to compare valued-added rankings to other measures of teacher quality, and to the extent that informal comparisons are made (as in the *LA Times* article), they sometimes don’t agree with common sense.

None of this means that value-added models are worthless—they are not. But like all mathematical models, they need to be used with care and a full understanding of their limitations.

How Is VAM Used?

Many studies by reputable scholarly groups call for caution in using VAMs for high-stakes decisions about teachers.

A RAND research report: The estimates from VAM modeling of achievement will often be too imprecise to support some of the desired inferences [McCaffrey 2004, 96].

A policy paper from the Educational Testing Service’s Policy Information Center: 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 [Braun 2005, 17].

A report from a workshop of the National Academy of Education:

Value-added methods involve complex statistical models applied to test data of varying quality. Accordingly, there are many technical challenges to ascertaining the degree to which the output of these models provides the desired estimates [Braun 2010].

And yet here is the *LA Times*, publishing value-added scores for individual teachers by name and bragging that even teachers who were considered first-rate turn out to be “at the bottom”. In an episode reminiscent of the Cultural Revolution, the *LA Times* reporters confront a teacher who “was surprised and disappointed by her [value-added] results, adding that her students did well on periodic assessments and that parents seemed well-satisfied” [Felch 2010]. The teacher is made to think about why she did poorly and eventually, with the reporter’s help, she understands that she fails to challenge her students sufficiently. In spite of parents describing her as “amazing” and the principal calling her one of the “most effective” teachers in the school, she will have to change. She recants: “If my student test scores show I’m an ineffective teacher, I’d like to know what contributes to it. What do I need to do to bring my average up?”

Making policy decisions on the basis of value-added models has the potential to do even more harm than browbeating teachers. If we decide whether alternative certification is better than regular certification, whether nationally board certified teachers are better than randomly selected ones, whether small schools are better than large, or whether a new curriculum is better than an old by using a flawed measure of success, we almost surely will end up making bad decisions that affect education for decades to come.

This is insidious because, while people debate the use of value-added scores to judge teachers, almost no one questions the use of test scores and value-added models to judge policy. Even people who point out the limitations of VAM appear to be willing to use “student achievement” in the form of value-added scores to make such judgments. People recognize that tests are an imperfect measure of educational success, but when sophisticated mathematics is applied, they believe the imperfections go away by some mathematical magic. But this is not magic. What really happens is that the mathematics is used to disguise the problems and

intimidate people into ignoring them—a modern, mathematical version of the Emperor's New Clothes.

What Should Mathematicians Do?

The concerns raised about value-added models ought to give everyone pause, and ordinarily they would lead to a thoughtful conversation about the proper use of VAM. Unfortunately, VAM proponents and politicians have framed the discussion as a battle between teacher unions and the public.

Shouldn't teachers be accountable? Shouldn't we rid ourselves of those who are incompetent? Shouldn't we put our students first and stop worrying about teacher sensibilities? And most importantly, shouldn't we be driven by the data?

This line of reasoning is illustrated by a recent fatuous report from the Brookings Institute, *"Evaluating Teachers: The Important Role of Value-Added"* [Glazerman 2010], which dismisses the many cautions found in all the papers mentioned above, not by refuting them but by asserting their unimportance. The authors of the Brookings paper agree that value-added scores of teachers are unstable (that is, not highly correlated year to year) but go on to assert:

"The use of imprecise measures to make high-stakes decisions that place societal or institutional interests above those of individuals is widespread and accepted in fields outside of teaching [Glazerman 2010, 7]."

To illustrate this point, they use examples such as the correlation of SAT scores with college success or the year-by-year correlation of leaders in real estate sales. They conclude that "a performance measure needs to be good, not perfect". (And as usual, on page 11 they caution not to use value-added measures alone when making decisions, while on page 9 they advocate doing precisely that.)

Why must we use value-added even with its imperfections? Aside from making the unsupported claim (in the very last sentence) that "it predicts more about what students will learn ... than any other source of

information,” the only apparent reason for its superiority is that value-added is based on data. Here is mathematical intimidation in its purest form—in this case, in the hands of economists, sociologists, and education policy experts.

Of course we should hold teachers accountable, but this does not mean we have to pretend that mathematical models can do something they cannot. Of course we should rid our schools of incompetent teachers, but value-added models are an exceedingly blunt tool for this purpose. In any case, we ought to expect more from our teachers than what value-added attempts to measure.

A number of people and organizations are seeking better ways to evaluate teacher performance in new ways that focus on measuring much more than test scores. (See, for example, the Measures of Effective Teaching project run by the Gates Foundation.) Shouldn't we try to measure long-term student achievement, not merely short-term gains? Shouldn't we focus on how well students are prepared to learn in the future, not merely what they learned in the past year? Shouldn't we try to distinguish teachers who inspire their students, not merely the ones who are competent?

When we accept value-added as an “imperfect” substitute for all these things because it is conveniently at hand, we are not raising our expectations of teachers, we are lowering them. And if we drive away the best teachers by using a flawed process, are we really putting our students first?

Whether naïfs or experts, mathematicians need to confront people who misuse their subject to intimidate others into accepting conclusions simply because they are based on some mathematics. Unlike many policy makers, mathematicians are not bamboozled by the theory behind VAM, and they need to speak out forcefully. Mathematical models have limitations. They do not by themselves convey authority for their conclusions. They are tools, not magic. And using the mathematics to intimidate — to preempt debate about the goals of education and measures of success — is harmful not only to education but to mathematics itself.

References

Audrey Amrein-Beardsley, *Methodological concerns about the education value-added assessment system*, Educational Researcher 37 (2008), 65–75. <http://dx.doi.org/10.3102/0013189X08316420>

Eva L. Baker, Paul E. Barton, Linda Darling-Hammond, Edward Haertel, Hellen F. Ladd, Robert L. Linn, Diane Ravitch, Richard Rothstein, Richard J. Shavelson, and Lorrie A. Shepard, *Problems with the Use of Student Test Scores to Evaluate Teachers*, Economic Policy Institute Briefing Paper#278, August 29, 2010, Washington, DC. <http://www.epi.org/publications/entry/bp278>

Henry Braun, *Using Student Progress to Evaluate Teachers:*

A Primer on Value-Added Models, Educational Testing Service Policy Perspective, Princeton, NJ, 2005. <http://www.ets.org/Media/Research/pdf/PICVAM.pdf>

Henry Braun, Naomi Chudowsky, and Judith Koenig, eds., *Getting Value Out of Value-Added: Report of a Workshop, Committee on Value-Added Methodology for Instructional Improvement, Program Evaluation, and Accountability*, National Research Council, Washington, DC, 2010.

<http://www.nap.edu/catalog/12820.html>

Donald T. Campbell, *Assessing the Impact of Planned Social Change*, Dartmouth College, Occasional Paper Series, #8, 1976. <http://www.eric.ed.gov/PDFS/ED303512.pdf>

Jason Felch, Jason Song, and Doug Smith, *Who's teaching L.A.'s kids?*, Los Angeles Times, August 14, 2010. <http://www.latimes.com/news/local/la-me-teachers-value-20100815.0.2695044.story>

Trip Gabriel, *Under pressure, teachers tamper with tests*, New York Times, June 11, 2010.

<http://www.nytimes.com/2010/06/11/education/11cheat.html>

Steven Glazerman, Susanna Loeb, Dan Goldhaber, Douglas Staiger, Stephen Raudenbush, Grover Whitehurst, *Evaluating Teachers: The Important Role of Value-Added*, Brown Center on Education Policy at Brookings, 2010. http://www.brookings.edu/reports/2010/1117_evaluating_teachers.aspx

Ted Hershberg, Virginia Adams Simon and Barbara Lea Kruger, *The revelations of value-added: An assessment model that measures student growth in ways that NCLB fails to do*, *The School Administrator*, December 2004.

<http://www.aasa.org/SchoolAdministratorArticle.aspx?id=9466>

David Hill, *He's got your number*, *Teacher Magazine*, May 2000 11(8), 42–47.

<http://www.edweek.org/tm/articles/2000/05/01/08sanders.h11.html>

Cory Koedel and Julian R. Betts, *Re-Examining the Role of Teacher Quality in the Educational Production Function*, Working Paper #2007-03, National Center on Performance Initiatives, Nashville, TN, 2007.

http://economics.missouri.edu/working-papers/2007/wp0708_koedel.pdf

Daniel Koretz, *Measuring Up: What Educational Testing Really Tells Us*, Harvard University Press, Cambridge, Massachusetts, 2008.

E. F. Lindquist, *Preliminary considerations in objective test construction*, in *Educational Measurement* (E. F. Lindquist, ed.), American Council on Education, Washington DC, 1951.

J. R. Lockwood, Daniel McCaffrey, Laura S. Hamilton, Brian Stetcher, Vi-Nhuan Le, and Felipe Martinez, *The sensitivity of value-added teacher effect estimates to different mathematics achievement measures*, *Journal*

of Educational Measurement 44(1) (2007), 47–67. <http://dx.doi.org/10.1111/j.1745-3984.2007.00026.x>

Daniel F. McCaffrey, Daniel Koretz, J. R. Lockwood, and Laura S. Hamilton, *Evaluating Value-Added Models for Teacher Accountability*, RAND Corporation, Santa Monica, CA, 2003.

http://www.rand.org/pubs/monographs/2004/RAND_MG158.pdf

Daniel F. McCaffrey, J. R. Lockwood, Daniel Koretz, Thomas A. Louis, and Laura Hamilton, *Models for value-added modeling of teacher effects*, *Journal of Educational and Behavioral Statistics* 29(1), Spring 2004, 67-101.

http://www.rand.org/pubs/reprints/2005/RAND_RP1165.pdf

RAND Research Brief, *The Promise and Peril of Using Value-Added Modeling to Measure Teacher Effectiveness*, Santa Monica, CA, 2004. http://www.rand.org/pubs/research_briefs/RB9050/RAND_RB9050.pdf

William L. Sanders and Sandra P. Horn, *Educational Assessment Reassessed: The Usefulness of Standardized and Alternative Measures of Student Achievement as Indicators of the Assessment of Educational Outcomes*, *Education Policy Analysis Archives*, March 3(6) (1995). <http://epaa.asu.edu/ojs/article/view/649>

W. Sanders, A. Saxton, and B. Horn, *The Tennessee value-added assessment system: A quantitative outcomes-based approach to educational assessment*, in *Grading Teachers, Grading Schools: Is Student Achievement a Valid Evaluational Measure?* (J. Millman, ed.), Corwin Press, Inc., Thousand Oaks, CA, 1997, pp 137–162.

Follow *The Answer Sheet* every day by bookmarking <http://www.washingtonpost.com/blogs/answer-sheet>. And for admissions advice, college news and links to campus papers, please check out our **Higher Education** page. Bookmark it!