FairTest

National Center for Fair & Open Testing

Student Test Scores: An Inaccurate Way to Judge Teachers

By Monty Neill[1]

Many states are considering how to evaluate teachers and administrators in response to the demands of the federal Race to the Top program. RTTT requires states to include student standardized test scores as a "significant" factor in that evaluation. States (and districts) are considering or already implementing what are generally termed "valued added measurement" models (VAM), which are complex statistical procedures used to analyze changes in students' test scores.

Unfortunately, VAM is a bad tool for making judgments about teachers or administrators, as extensive research shows VAM will produce inaccurate and unfair judgements. Many good teachers could receive bad rankings, and vice versa. At the same time, this use of standardized tests will only intensify the control of testing over curriculum and instruction.

U.S. Education Secretary Arne Duncan, the Gates Foundation, business groups and leading newspapers are pushing VAM hard. Some states have decided to make it count for *half* a teacher's evaluation — for those who teach tested subjects.

So what is VAM — and why is it dangerous? I will go into some depth on some (but not all) of the key reasons.[2]

What is "Value-Added Measurement"? VAM attempts to compare test score changes among different teachers' students. As students move from grade 3 to 4 to 5, etc., their scores are tracked. With this data, a state or district tries to measure how much "value" (test score gain) a given teacher provides to her students each year. More precisely, the point is to consider whether a student made more, the same or less gain than his/her peers, then to decide how much of that gain is due to the teacher's efforts. The results are used to rank teachers. In some states, the rankings have been publicized in the newspapers.

To accomplish this, states rely on complex statistical formulas. They do so because states and districts are trying to figure out how much of relative gain or loss is due to the teacher rather than

[1] An earlier version of this paper was published in the Massachusetts Citizens for Public Schools PS *Backpack*, Spring 2011.

[2] For a short op ed that lists seven reasons to not use VAM for judging educators and explains the dangerous consequences of doing so, link to the B. Baker (March 2011) article cited in the references.

other factors. For multiple reasons, these statistics too inaccurate for use in making high-stakes decisions about educators. But there are other fundamental flaws, starting with the tests themselves.

VAM is limited by the quality of the tests on which it is based. VAM is no better than the tests on which it is based; it is in fact, simply a different way to use the same old tests. Unfortunately, standardized tests are extremely narrow and inadequate measures of student learning. The gains measured by VAM leave out a whole range of important knowledge and skills that are not assessed by any state's standardized math and English tests. VAM creates additional distortions and inaccuracies. If used to make important decisions about teachers or principals, it will intensify the already damaging control tests have over teaching and learning (see, in general, materials at http://www.fairtest.org).

Most states test only reading and math, meaning that VAM data is only available for a limited number of teachers (some studies say only one in five). Most teachers could not be judged by VAM results. To "solve" this problem, states and districts are proposing a huge expansion in the amount of testing, so every teacher will have her "own" VAM scores. For example, Charlotte-Mecklenberg, NC, recently decided to spend $1.9 million to create 52 new tests. Florida is moving in the same direction, despite a huge outcry that led then-Governor Christ to veto a legislative proposal to make dozens more tests. Inevitably, these tests are all or predominantly multiple-choice and fail to assess higher order thinking. They also tend to impose a straitjacket on teachers, curriculum and instruction.

VAM has numerous technical flaws that undermine its accuracy. On the surface, it seems to make sense to look at student gains, rather than students' one-time scores. Progress is important, and single-shot tests are clearly dependent on a student's class, race, disability status and knowledge of English. VAM promises to take account of students' backgrounds. But, research shows, the statistical techniques do not adequately adjust for different student populations, so teachers are still judged based on their students' background status.

Fundamentally, VAM models are unable to demonstrate a causal link between a teacher and changes in student test scores. One reason is that students are not randomly assigned. Since districts, schools and teachers have students with varying background characteristics that affect student learning and progress, VAM would have to incorporate statistical techniques that can adequately adjust for non-random assignment. Otherwise, one cannot presume that different outcomes are due to the teachers. But VAM does not adequately do so (c.f., Baker 2011, March). For example, prominent measurement expert Howard Wainer explains that issues such as the influence of other students in a classroom (peer effects) are among many that undermine causal claims. Peer effects were prominent in a recent Montgomery County, MD, study (Kahlenberg 2010).

While VAM attempts to rule out the consequences of both student background factors and other in-school factors, it can only do so in a very partial way. Thus, teachers are judged by VAM in large part for things over which they have no control. This shows up, in part, in very erratic results.

For example, the researchers Peter Schochet and Hanley Chiang show that, even with three years of student test scores, teachers are rated inaccurately one time out of four. It is worse with only one or two year's data. Tim Sass reports that more than two-thirds of the bottom-ranked teachers one year had moved out of the bottom ranks the next year. One third moved from the bottom 20 percent one year to the top 40 percent the next. Only a third who ranked highest one year kept their top ranking the next, and almost a third of the formerly top-ranked teachers landed in the bottom 40 percent in year two. Other studies have found similar instability in "value-added" rankings. Wayne Au explained, "Because of these error rates, a teacher's performance evaluation may pivot on what amounts to a statistical roll of the dice."

Gates Foundation-supported research argued that VAM based on state tests correlated with other measures, including different tests that supposedly measured deeper, conceptual understanding. But noted economist Jesse Rothstein reported the actual results of the Gates-funded study don't support that conclusion. The correlation between the two kinds of tests is "only slightly better than coin tosses."

There are many other technical flaws in this measurement process. For example, when different VAM models are applied, the results can vary dramatically (Briggs and Domingue). The consequences is that they are too inaccurate to justify their use in making decisions about educators (as indicated by many of the cited articles; in general, see Au, and B. Baker's March 2011 op ed.). That's on top of the limits and flaws of the underlying tests. But, it gets worse.

The circular "logic" underlying VAM. Rothstein also showed that the Gates study only included other measures (such as observations) if those measures correlated positively with VAM results. That is, they decided ahead of time that VAM was best, then ignored anything that did not correspond with the VAM results. (See also DiCarlo, 3/2011.)

This sort of circular reasoning is common in the use of VAM for evaluating teachers. VAM use assumes that those teachers whose students make higher test-score gains are better teachers. When VAM scores are used to evaluate teachers, those whose students show higher rates of test score gains are identified as better. Those whose students gain at lower rates are deemed ineffective. This is a flaw, for example, in the paper by Goldhaber and Theobald that claimed VAM was a superior method for determining which teachers to lay off: the assumption guaranteed the finding.

It is a mistake to draw conclusions from VAM data about teacher "effectiveness." To start, VAM makes unwarranted assumptions student learning trajectories. As Brookline, Mass., parent Lisa Guisbond explained,

> Expecting good teachers to "*routinely* impart a year-and-a-half-gain in student achievement" in one year is like expecting the housing bubble to inflate indefinitely. This proved impossible for the housing market, and it's impossible for human beings…
>
> One problem is, and educators know this from working with actual children, children do not develop and learn on a steady upward curve, no matter how stupendous a teacher they have. My own kids have had some extraordinary (award-winning) teachers

> and have not even made a year's worth of gain on their watch, based on their developmental timetable and readiness to learn.
>
> In real life, an experienced teacher may lay the foundation for a big leap a year or two later. This may happen on the watch of a lesser teacher, who happened to be around to reap the benefits. Who should get the credit?"

The "year-and-a-half gain" claim that has been bandied about the VAM supporters is in fact based on extrapolations done by Eric Hanushek, not on any actual evidence from live children in with actual teachers. Similarly, a paper by William Sanders and June Rivers relied on statistical projections for their claim that if only a low-scoring student had X number of "superior" teachers in a row, the student would close the achievement gap. Writ large, that means if low-income or racial minority groups that commonly score lower only had better teachers, they would close the income/race score gaps (DiCarlo, 2011, March).

As Mathew DiCarlo pointed out, researchers "took the average one-year gains among students of "top teachers" (however defined), and then determined how many of these one-year gains are equivalent to the average aggregate achievement gap… [O]ne must be very careful in applying the estimated one-year testing gains among a large, diverse group of students to a hypothetical scenario in which a specific "type" of student (e.g., low-income) moved from one specific score to another (e.g., moving from the average for free lunch-eligible students to that of non-eligible students) over a period of years." In addition, any one teacher's contribution to student scores "decays" rather quickly over time; it has limited persistent effect.

DiCarlo then points out the fact (noted above) that teachers cannot reliably be identified as superior (even at raising test scores, never mind the real learning the scores purport to measure). As a result, districts could not even reliably determine who would be able to produce those consistent test score gains.

There are many other flaws and weaknesses with VAM. For example, who is the teacher of record if a student moves during the year or leaves school for two months in the winter, or if there are multiple teachers for a course? Also, to work best, VAM has to assume a linear relationship among different topics in a subject, such as Algebra and Geometry in math, or Biology and Physics in science. But both subjects and learning are multi-dimensional, not linear (Bracey).

Conclusion. There are so many variables, so many things outside the control of schools and teachers that VAM cannot account for, so many statistical limits and flaws, that it is clearly unfair to use VAM as a tool for judging teachers. Some teachers would be unjustly fired, and many would quit. Not only would teachers be inaccurately and unfairly judged, they would feel pressured to teach even more intensely to the test. That would further damage and limit our children's education.

To argue that this flawed measurement tool would be only one of a number of ways of evaluating teachers does not address the problems it would cause. It is not needed as part of a high-quality evaluation system for educators.

Simple indicators to evaluate or pay employees are used only rarely in other professions (Adams, et al.; FairTest). When they are, the professionals engage in their version of teaching to the test, with often-disastrous results. Wall Street paid its speculators based on simplistic measures, and we are still suffering the consequences.

Bruce Baker (2011, March) summarized the research evidence: VAM "just doesn't work, at least not well enough to even begin considering using it for making high-stakes decisions about teacher tenure, dismissal or compensation... In fact, it will likely make things much worse. Establishing a system where achieving tenure or getting a raise becomes a roll of the dice and where a teacher's career can be ended by a roll of the dice is no way to improve the teacher work force."

Teachers deserve high-quality evaluation, for fairness and to help them improve. MCAS is too weak to use for making decisions about students. When all the limitations and errors of VAM are factored in, it renders the process "valueless addition" for teachers and their students.

Bibliography

Au, Wayne. 2010-11. "Neither Fair Nor Accurate." In *Rethinking Schools*, Winter, pp. 34-38. Available at http://www.rethinkingschools.org/archive/25_02/25_02_au.shtml.

Baker, Bruce. 2011, Feb. 16. "Reformy Disconnect: 'Quality Based' RIF?" http://schoolfinance101.wordpress.com/2011/02/16/reformy-disconnect-quality-based-rif/

Baker, Bruce. 2011, March 13. "7 reasons why teacher evaluations won't work." *The Record.* http://www.northjersey.com/news/education/evaluation_031311.html?page=all. An excellent, popular summary of key flaws in the use of VAM to judge teachers.

Baker, Eva, et al. 2010. August. *Problems with the Use of Student Test Scores to Evaluate Teachers*. Economic Policy Institute. http://www.epi.org/publications/entry/bp278. Finds that "If the quality, coverage, and design of standardized tests were to improve, some concerns would be addressed, but the serious problems of attribution and nonrandom assignment of students, as well as the practical problems described above, would still argue for serious limits on the use of test scores for teacher evaluation."

Bracey, Gerald R. 2007, July. "Evaluating Value-Added." *FairTest Examiner*.

Briggs, Derek, and Ben Domingue. 2011, 2. "Due Diligence and the Evaluation of Teachers: A Review of the Value-Added Analysis Underlying the Effectiveness Rankings of Los Angeles Unified School District Teachers by The Los Angeles Times." National Education Policy Center, School of Education, University of Colorado at Boulder. http://nepc.colorado.edu/publication/due-diligence.

Corcoran, Sean P. 2010. "Can Teachers be Evaluated by their Students' Test Scores? Should They Be? The Use of Value-Added Measures of Teacher Effectiveness in Policy and Practice. Annenberg Institute for School Reform."

http://www.annenberginstitute.org/products/Corcoran.php. Finds that "the promise that value-added systems can provide such a precise, meaningful, and comprehensive picture is not supported by the data" (p. 28).

Di Carlo, Matthew. 2011, January 14. "The biggest flaw in Gates value-added study." *The Washington Post, Answer Sheet.* http://voices.washingtonpost.com/answer-sheet/guest-bloggers/the-biggest-flaw-in-the-gates.html. In this blog, he lucidly explains some of J. Rothstein's findings about the Gates-funded report from Kane.

DiCarlo, Matthew. 2011, March 31. "The nonsense behind the 'X consecutive teachers' argument." *The Washington Post, Answer Sheet.* http://www.washingtonpost.com/blogs/answer-sheet/post/the-nonsense-behind-the-x-consecutive-teachers-argument/2011/03/29/AFlU345B_blog.html. Concludes that VAM cannot be used to accurately determine who are better teachers, and should not be used for making decisions about teachers.

FairTest. 2009, November. "Paying Teachers for Student Test Scores Damages Schools and Undermines Learning." http://www.fairtest.org/paying-for-student-test-scores-damages-schools. Summarizes national and international research findings, in education and elsewhere, with a bibliography.

Gates Foundation. 2010. *Learning about Teaching: Initial Findings from the Measures of Effective Teaching Project*. MET Project Research Paper. Seattle, Washington: Bill & Melinda Gates Foundation. http://www.metproject.org/downloads/Preliminary_Findings-Research_Paper.pdf. Argues that VAM correlates well with measures of higher-order thinking, a point strongly rebutted by J. Rothstein (2011), who said that the MET's own "data in fact indicate that a teachers' value-added for the state test is not strongly related to her effectiveness in a broader sense" and ""do not support the conclusions drawn from them."

Goldhaber, Dan, and Roddy Theobald. 2010. "Assessing the Determinants and Implications of Teacher Layoffs." Center for Education Data & Research, University of Washington Bothell, CEDR Working Paper 2010-07. Makes claim that VAM is better tool for making layoff determinations than is seniority; assumptions rebutted by DiCarlo (3/11) and B. Baker (2011).

Guisbond, Lisa. 2011, April 27. Testimony to the Mass. Board of Elementary and Secondary Education.

Kahlenberg, Richard D. 2010, Oct. 15. "Housing Policy Is School Policy." *Education Week.* http://www.edweek.org/ew/articles/2010/10/20/08kahlenberg_ep.h30.html?qs=Montgomery+County+Maryland

McCaffrey, D., Koretz, D., Lockwood, J.R., and Hamilton, L. 2005. "Evaluating Value-Added Models for Teacher Accountability." Santa Monica: RAND Corporation. Concluded that "the research base is currently insufficient to support the use of [value-added methods] for high-stakes decisions about individual teachers or schools."

National Research Council, Board on Testing and Assessment. 2009. "Letter Report to the U.S. Department of Education on the Race to the Top Fund." National Academy of Sciences, available at http://www.nap.edu/catalog.php?record_id=12780. Recommended against the requirement in Race to the Top to include student test scores in the evaluation of teachers.

Rothstein, Jesse. 2011, January. "Review of 'Learning About Teaching.'" National Education Policy Center. http://nepc.colorado.edu/thinktank/review-learning-about-teaching. Important critique of flaws in Kane's Gates-funded study, with generalizable conclusions about VAM.

Sass, Tim R. 2008, November. "The Stability of Value-Added Measures of Teacher Quality and Implications for Teacher Compensation Policy." National Center for Analysis of Longitudinal Data in Education, Policy Brief 4. http://www.urban.org/UploadedPDF/1001266_stabilityofvalue.pdf. (The full paper on which the brief is based is at http://www.urban.org/UploadedPDF/1001469-calder-working-paper-52.pdf.)

Schochet, Peter Z., and Hanley S. Chiang. 2010, July. "Error Rates in Measuring Teacher and School Performance Based on Student Test Score Gains." U.S. Department of Education, Institute for Education Sciences. NCEE 2010-4004. http://ies.ed.gov/pubsearch/pubsinfo.asp?pubid=NCEE20104004.

Schwartz, Heather. 2010. "Housing Policy Is School Policy: Economically Integrative Housing Promotes Academic Success in Montgomery County, Maryland." New York: Century Foundation. http://tcf.org/publications/pdfs/housing-policy-is-school-policy-pdf/Schwartz.pdf. Demonstrates power of peer effects, value of housing desegregation.

Wainer, Howard. 2011, February. "Value-Added Models to Evaluate Teachers: A Cry For Help." *Chance.* http://chance.amstat.org/2011/02/value-added-models/. Explains three major flaws in VAM.