FairTest

National Center for Fair & Open Testing

Testimony in Response to Proposed Regulations on Evaluation of Educators, 603 CMR 35.00

From: Monty Neill, Ed.D., Executive Director

To: Massachusetts Board of Elementary and Secondary Education

Date: April 27, 2011

I respectfully request that the Board of Elementary and Secondary Education reject and send back to the Department the draft regulations on teacher evaluation (603 CMR 35.00). I ask that you do this today – this draft is not one the Board should submit to the public for comment. Many changes must be made first.

The draft is so flawed its implementation will not only put many educators unfairly at risk, but it will also intensify teaching to flawed and limited standardized tests. That, in turn, will exacerbate, not ameliorate, the gaps in opportunity to learn and educational outcomes, by race and class, that plague the Commonwealth. It will frequently act to undermine, not promote, teacher and education quality, though the vagueness and flexibility in the document means the damage probably will not occur in some districts. Such districts are more likely to be wealthy, and thus the flexibility is likely to further privilege those educators and students who are already most advantaged.

If the Board does not on April 27 return the draft Regulations to the Department, and barring fundamental positive changes made in response to public comment and the Board’s own thinking, the BESE should reject these when public comment is completed. Should the Department make extensive revisions in response to public comment, I hope and expect that those revisions will themselves be submitted for public comment before any vote for approval by the Board. This could delay the final vote, but it is far more important to have a healthy, educationally beneficial evaluation system that minimizes harmful consequences, rather than to enact regulations hastily.

In this testimony, I will explain a series of problems with the draft regulations. At the end, I append a more thorough explanation of the flaws, limitations and dangers of using the “growth” version of student MCAS results to judge individual teachers. The ten issues I will address below are:

1. Too many new tests.
2. Fostering an overemphasis on test preparation.
3. Misuse of “growth” models.
4. Unreasonably counting student “learning gains” separately and twice.
5. No meaningful consideration of classroom and school evidence of student learning.
6. Unclear expectations for “student learning” gains.
7. Probable highly different applications of the Regulations in different districts, to the disadvantage of students in heavily low-income communities.
8. Potential over-placement of teachers in “directed-growth” plans who will be at risk of termination “at any time.”
9. Pitting teachers against one another.
10. Damage to school climate and teacher morale.

1. Too many new tests. The draft regulations will require districts to make or purchase dozens of new tests. This expensive undertaking will come at a time of teacher layoffs and extensive cutbacks in schools across the Commonwealth. Many districts lack the capacity to develop high quality assessments to use across classrooms and schools. They are therefore apt to buy commercial products that do not reflect Massachusetts standards and will tend to reduce teaching and learning to what can be measured by multiple-choice questions focused on rote learning. This will dumb down the quality of curriculum and instruction, guaranteeing our students will be less prepared to be effective citizens, successful in college or at well-paying jobs, and competent life-long learners. This damage will alienate many students who will, quite reasonably, reject wasting their time and attention on drill and kill test preparation.

In the best of circumstances, districts would work with their teachers to create and score high-quality performance assessments. But few districts will be able to do so, particularly with their now diminished resources. This is especially true for low-wealth districts, which are therefore the ones most likely to purchase weak commercial products. This will exacerbate, not narrow, the opportunity to learn gap that now exists by race and class.

Alternatively, the state will devote its scarce resources to crafting dozens of new tests. But, we already know that current MCAS tests are not adequate measures of what students must be able to know and do for future success. This is true even if one believes that MCAS is better than other state tests. (My comparative review finds, for example, the MCAS language arts tests are very similar in most respects to those of Texas, which has never been identified as having high-quality tests.)

Achieve, a group that supports high-stakes testing, has identified attributes sought by college instructors of first-year students as well as those sought by higher-paying employers. It is very clear that most of those attributes are not measured by MCAS. There is no reason to believe the comprehensive and higher-order skills sought by colleges and employers will be measured by new commercial or Department-made tests.

In short, the entire proposal is predicated on the false assumption that districts can create high-quality standardized assessments or that they have the will and resources to quickly develop a wide range of portfolios of student work that can be used across entire districts. The reality is this is nearly certain not to happen, and the results will cause serious damage to education.

2. Fostering an overemphasis on test preparation. Inclusion of MCAS and additional similar or worse tests, coupled with high-stakes for teachers and administrators, will intensify the reduction of instruction to test preparation. Numerous studies, including one just released in Baltimore (Plank and Condliffe), find that the emphasis on testing reduces intellectual and academic quality in classrooms. Indeed, this has been a widespread complaint about the impact of the federal No Child Left Behind Act (NCLB). (See Neill, Guisbond and Schaeffer for more evidence on this point.) The highly-respected Phi Delta Kappa/Gallup annual surveys of education have documented this concern among parents and the public, while numerous studies have found teachers have expressed concern over and opposition to the pressure to focus on test scores.

Thus, an effort that purports to evaluate educators in order to improve the quality of teaching and administration will rather certainly undermine the actual quality of instruction and learning.

3. Misuse of "growth" models. The draft regulations mandate the use of so-called "growth" measures on existing MCAS tests to rate teachers. The state's "Student Percentile Growth ," or SPG, is a variant on what has been misleadingly labeled "value added" measurement or VAM. A good deal of research has been conducted on VAM, and I will, below, explore in detail the issue of using VAM in any form, including SPG. Here, I will briefly describe some of the significant flaws:

Much is uncertain in the regulations. It appears for example that student test score gains for the previous year will be used to judge a teacher. It may be that each year, multiple years of test score gains will be used. Maybe this will be left up to districts or forthcoming guidance from the Department. I any event, there will be a great deal of measurement error, so much that it will be thoroughly unfair to subject educators to the whims of statistical chance. This issue is compounded by the requirement in the draft that teachers identified as producing low student gains have only one year to improve. Ironically, the yearly fluctuations in growth data could help some teachers "improve" in year 2 compared with year 1, but there remains far too great a chance of mislabeling teachers.

Growth/VAM models such as SPG often incorrectly claim that the test-score-gain measures can effectively ensure that students socio-economic status, family and community status, disability status or English-language learner status, do not affect students' annual test-score gains. But they do, and it has been well-demonstrated that, in fact, students from more privileged backgrounds, who are native English speakers, who do not have a disability, possess advantages that do not end when they enter the classroom, but carry over each day in the school year and during the summer. Thus, for example, low-income students suffer summer learning loss, while high-income students actual show gains on test scores when not in school over the summer. (A dangerous and expensive "solution" to this problem is to have more tests, in fall and spring, so only gains during the school year are measured. This would limit the impact of summer learning loss or gain, but not fully account for it, as those who gained more in the summer have, in effect, more learning capital to apply to even faster gains during the school year.) The state's procedure of comparing students with "similar score histories" can partially mitigate these effects, but the research on various VAM models finds that such efforts to control for past results are only partially successful.

In addition, students are not randomly assigned to teachers or to schools. Indeed, poverty and racial isolation mark many schools and even whole districts, while most schools and districts engage in some form of tracking that effectively gives some teachers students who start with significant advantages. A recent study from Montgomery County, Maryland, found that low-income/racial minority students who attended schools with predominantly wealthier students made faster gains than comparable students who attended schools that had been provided extra resources because of their preponderance of lower-income students (Kahlenberg). This is likely in large part a peer-effect, one that will play out in many schools in the Commonwealth. (Peer effects have been noted in other studies, as well.)

Thus, teachers in schools with mixed SES characteristics will benefit or suffer based on student assignments and teacher's own class assignments, while educators in whole schools and districts will suffer or benefit based on existing social inequalities that growth/VAM does not adequately account for. Sorting students' scores based on their score histories, as SPG does, will in no way solve this problem. In the Montgomery County case, teachers whose students benefited from peer effects would be rewarded, while teachers whose students did not so benefit would be punished.

Also, I find the regulations unclear on just how students will be compared. The 2009 SPG report from the Department mentions in passing that one can compare, for example, students with disabilities, but the draft regulations do not say whether that will be expected, allowed or disallowed as districts apply the regulations to their educators. Clearly, such comparison groupings should be expected (which would not preclude other forms of comparisons), though in the case of SPED and ELL, by no means are all disabilities the same, nor students with similar disabilities or level of English evenly distributed – again, raising questions about the accuracy of SPG for evaluating teachers and administrators.

4. Unreasonably counting student "learning gains" separately and twice. The matrix provided in the Commissioner's memorandum accompanying the draft regulations makes clear that "Measures of Student Performance" (probably just test scores in most districts) will be counted twice when evaluating educators. In the matrix, the two (or more) components of "performance" constitute the horizontal component. "Measures of Educator Practice" form the vertical component, but "Educator Practice" must itself include "student performance." Thus, as it combines the two measures, the matrix not only counts "Student Performance" by itself as much as "Educator Practice," but test scores are to form a significant (though unspecified) portion of "Educator Practice."

The regulations should make clear that student learning gains are to be considered when evaluating "Educator Practice" but not separately, as is done in the memo. I partially address the what and how of this elsewhere in these comments, but my purpose here is not to propose an alternative for the state to consider, but to analyze the flaws and dangers of the draft Regulations.

[Note: The draft regulations at 35.08 read: "(5) At a minimum, multiple measures of student learning, growth, and achievement shall be used in rating the Curriculum, Planning, and Assessment and Teaching all Students standards for teachers and the Curriculum, Instruction, and Assessment and Management and Operations standards for administrators." However, the discussion at 35.03(1), "Standards and Indicators of Effective Teaching," does not require inclusion of evidence of student learning (test scores). This creates some ambiguity, at best, in the draft regulations.]

5. No meaningful consideration of classroom and school evidence of student learning. The Task Force emphasized classroom and school measures determined by the teacher and the evaluator. These are essentially ignored in the draft Regulations. Even the "additional measures" must be "comparable across grades and subject matter district-wide as determined by the superintendent and approved by the Department" (p. 17). This seems to essentially eliminate an important area. The Board should insist that in this case, the Regulations be changed to accord with the Task Force report, and significant weight be given to evidence of learning developed in relation to a teacher's own curriculum and a school's own goals.

6. Unclear expectations for "student learning" gains. The Commissioner's Memorandum accompanying the draft Regulations states, at "Student Performance Measures" (p. 7):

- Evaluators determine whether each educator's impact on student learning is low, moderate, or high. For each year of instruction: moderate impact is represented by student learning gains of a year's growth; growth of less than one year represents low impact; and high impact is represented by growth of more than one year. As with expected MCAS growth, it will be important for districts to clearly identify what constitutes low, moderate, and high student learning growth based on guidelines that the state will develop.

I did not find the "definition" of moderate, low and high impact in the actual draft Regulations, so their import is not clear. If a year's growth means the average or median rate of growth for students (given how SPG is determined based on test-score history), then it is certainly possible that something close to half the state's teachers each year will by this standard be deemed failures. Maybe the state will ease the problem by defining "average" as a range that includes those only somewhat below average (as it does in SPG itself), but this will not be sufficient to prevent many teachers every year from being unfairly labeled. In any event, it appears that the Department is applying a norm-based rather than standards-based approach as it incorporates SPG into the evaluation process. This could be countered by the use of the other assessments, but a goal here appears to identify some (possibly large) class of teachers as inadequate.

Of course, given the role of evaluator judgment and the use of test/assessment data for which SPG does not exist, it is very possible that districts will rationally decide not to fail half their teacher annually. But some districts may choose to do so, for among other things, it could provide them with enormous flexibility in whom to lay off or fire. (This is very clearly the case with the Washington, DC., "evaluation" system negotiated by Superintendent Rhee with the teachers union in exchange for significant pay increases.) Thus, a politically unpopular, educationally controversial, union-supporting or other teacher whose SPG average is a bit low, even if due to statistical error, is on one-year notice and may in fact be fired at any time. (see, for example, 35.06 (7)(b)2). Thus, some teachers will face a rigged game, one in which their fates will be determined in large part by student, school and community characteristics over which they have no control.

This SPG requirement in the regulations will put teachers whose students are statistically likely to make slower than average gains, regardless of the quality of teaching, at serious risk of losing their jobs. Under those circumstances, teachers will rationally seek to work with students who are more likely to make average or better score gains. They will seek to work in wealthier communities, where for the most part their working conditions and salaries are already superior. They will seek to avoid students in their schools who are likely to gain less. In these circumstances, it is likely that the weaker teachers will in fact end up in classrooms with the most needy students. It will cause teachers to look anxiously and fearfully at some of their students. It will encourage them to engage in disciplinary actions that will push some students out of their classrooms and schools.

7. Probable highly different applications of the Regulations in different districts, to the disadvantage of students in heavily low-income communities.

It is reasonable and necessary that expert judgment determine an educator's summative status. Such judgments cannot be rendered rationally by use of a statistical formula that inevitably includes only a few factors and fails to consider the overall situation in which an educator works. However, it is all too likely that some districts will choose to weight SPG and other standardized test scores very highly. (Boston has already signaled its intention to do so.) In such districts, teaching to the standardized tests, MCAS and others, is likely to intensify even beyond what it now is. Because such tests, including MCAS, are unable to adequately measure many important areas of desired learning (as noted above at point 2), students in these districts will become even less likely to acquire the knowledge and skills needed for adult success. In these circumstances, professional judgment will be functionally reduced in favor of statistical procedures based on flawed and limited tools.

(I am not unaware that the flexibility and evaluator judgment also can be used arbitrarily and unfairly against administrators or teachers whom the higher-ups seek to remove. Perhaps unions will remain able to ensure some protections for teachers in the development of district procedures. Neither the draft Regulations nor the Commissioners Memorandum seem to address what happens if unions and administrators/school committees reach an impasse.)

8. Potential over-placement of teachers in "directed-growth" plans who will be at risk of termination "at any time." According to the draft Regulations (at 35.06 (7)(b)2), the evaluator "shall" place teachers with "professional status" who are rated "proficient" or even "exemplary" in "directed growth" plans if their student "learning gains" are judged to be low. Such teachers may be "dismissed at any time." That is, a teacher who is judged to be "exemplary" by a presumably competent evaluator may nonetheless be put at risk of being fired "at any time" because her/his students' scores on the assessments are insufficient. The issue of statistical inaccuracy comes into play here. But there is also something bizarre in the notion that an "exemplary" or even a "proficient" teacher must be in a "directed growth" plan and put at risk of being fired. (Ironically, for teachers who have not attained "professional" status, being placed in this group is at the discretion of the evaluator and not mandated.) This requirement certainly runs counter to the Commissioner's claim (p. 6) that "evaluation is primarily about development and not primarily about sorting and shedding."

In this case, at a minimum, any "proficient" or "exemplary" teacher should not be placed at risk of termination and should be in a "self-directed" plan (or as appropriate the "developing teacher/administrator plan”). (In the matrix in the Commissioner’s memo, on page 9, this would move them from the ‘green’ to the ‘yellow’ zone.)

9. Pitting teachers against one another. If enacted, this proposal will in some instances pit teachers against one another. This is because using the norm-referenced metric of “average” student learning gains, especially for SPG, means that if your students score better, I am automatically at greater risk of falling below average. Teachers would have a disincentive to help each other improve. This clearly runs counter to evidence showing teacher collaboration is very important for school climate and improvement, and for student opportunity to learn.

10. Damage to school climate and teacher morale. Because of the rather certain negative consequences in teacher to teacher and teacher to student relations, as well as a reduction in the quality of curriculum and instruction, school climate will inevitably worsen. This, in turn, will further increase the likelihood that good teachers and administrators will leave the profession, even if facing an unfair and irrational “evaluation” system does not already encourage them to leave. They especially will leave situations perceived as more unfair, more arbitrary, or harder in which to ensure students make the required test-score gains. These combined circumstances will certainly make teaching less attractive, particularly given the rather constant attacks on the profession waged by big business and the media and coupled with attacks on teachers’ pensions and health care.

One might wish teachers will not defend themselves against their students and fellow teachers, that they will not narrow and dumb-down their teaching to fit the tests that will rain down on them, but that is to wish that teachers don’t act as most humans will act when presented with sanctions and rewards by their bosses. For example, research has demonstrated that employees from doctors to bus drivers will alter their behaviors at the expense of patients or riders to protect themselves from accountability requirements (Adams, et al.; FairTest).

Conclusion: The draft Regulations combine flawed statistical procedures and inevitably-inadequate tests with punitive sanctions for teachers in ways that will seriously harm public education in the Commonwealth. The rhetoric in the Commissioner’s memorandum and the draft regulations drape the process in high-sounding, reasonable, even noble goals of improving the quality of teaching. Do not be misled. The actual mechanics of the draft regulations will produce the opposite.

Thus, the Board must reject these regulations and insist that the Department re-write the Regulations so that they do, in fact, support improved teacher and administrator quality and thus contribute to improved student learning.

A more detailed analysis of VAM follows the references for this section.

References

Note that I have only lightly referenced the points in these comments. I can provide further evidence on request.

Achieve. 2005, February. *Rising to the Challenge: Are High School Graduates Prepared For College And Work? A Study of Recent High School Graduates, College Instructors, and Employers.*

Adams, S.J., Heywood, J.S., and Rothstein, R. 2009. *Teachers, Performance Pay and Accountability: What Education Should Learn from Other Sectors.* Washington, DC: Economic Policy Institute.

Chester, Mitchell. 2011, April 16. Memorandum: Proposed Regulations on Evaluation of Educators. http://www.doe.mass.edu/boe/docs/0411/item1.pdf.

FairTest. 2009. Paying Teachers for Student Test Scores Damages Schools and Undermines Learning. http://fairtest.org/paying-for-student-test-scores-damages-schools.

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.

Massachusetts Department of Elementary and Secondary Education. 2011. Proposed Regulations on Evaluation of Educators. 603 CMR 35.00. http://www.doe.mass.edu/boe/docs/0411/item1_p603cmr35.pdf.

Neill, Monty, Lisa Guisbond and Bob Schaeffer. 2004, May. *Failing Our Children: How "No Child Left Behind" Undermines Quality and Equity in Education, and an Accountability Model that Supports School Improvement*. FairTest. http://fairtest.org/node/1778.

Plank, Stephen B., and Barbara Condliffe. 2011, February. *Pressures of the Season: A Descriptive Look at Classroom Quality in Second and Third Grade Classrooms.* Baltimore Education Research Consortium.

Continued on Next Page: Analysis of VAM

Student Test Scores: An Inaccurate Way to Judge Teachers

By Monty Neill

Massachusetts is considering how to evaluate teachers in response to the demands of the federal Race to the Top program. "Student Growth Profiles" (SPG) based on MCAS scores will be a significant part of teacher evaluations, according to a proposal before the state Board of Elementary and Secondary Education. SPG is one variant of what are generally termed "valued added measurement" models (VAM). Unfortunately, VAM is a bad tool for making judgments about teachers. Extensive research shows VAM will inaccurately and unfairly judge teachers. Many good teachers could receive bad rankings, and vice versa. At the same time, this use of MCAS will only intensify the control of testing over curriculum and instruction.

This is a national issue. 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.[1]

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 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).

[1] 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.

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.