Flawed Massachusetts Teacher Evaluation Proposal Risks Further Damage to Teaching and Learning

By the Massachusetts Working Group on Teacher Evaluation of the National Center for Fair & Open Testing

Contents

Lisa Guisbond and Jacqueline King, Co-Editors

Working Group Members
Lisa Guisbond, *Policy Analyst, FairTest*
James Horn, Ph.D., *Associate Professor, Education, Cambridge College*
Jacqueline King
Jonathan King, Ph.D., *Professor, Molecular Biology, MIT*
Louis Kruger, Psy.D., Director, *Northeastern University School Counseling Program*
Monty Neill, Ed.D., *Executive Director, FairTest*
Ann O'Halloran, *Angier School Teacher, retired, 2007 Massachusetts Preserve America History Teacher of the Year*
Bill Schechter, *Lincoln-Sudbury High School History Teacher, Retired*

FairTest

P.O. Box 300204
Jamaica Plain, MA 02130
(617) 477-9792
www.fairtest.org
Email: fairtest@fairtest.org
Find this report or the two page summary online at:
http://fairtest.org/flawed-ma-teacher-evaluation-proposal-report-home

Flawed Massachusetts Teacher Evaluation Proposal Risks Further Damage to Teaching and Learning

By the Massachusetts Working Group on Teacher Evaluation of the National Center for Fair and Open Testing

The Challenge: To Develop and Support Good Teaching, Keep Good Teachers in Classrooms

Knowledgeable, effective, caring and inspiring teachers are vitally important for high-quality student learning. A healthy education system has well trained, well supported, and properly compensated professionals. In countries with leading public education systems, such as Finland and Singapore, the community as a whole respects teachers. Among the most striking features of nations cited for outstanding academic outcomes are their professional recruitment, development and support practices, as well as the stature accorded teachers. Sadly, this sets them apart from our own country.

Massachusetts, along with the rest of the nation, faces a looming exodus of experienced teachers. In the next 10 years, more than 50% of the nation's 3.2 million public school teachers will become eligible for retirement. Our schools must also cope with a constant turnover of new teachers who leave the profession before they even begin to gain their footing. Nationwide, 46% of teachers quit before their fifth year. About 20% of teachers in urban districts leave every year. While it is common to cite the difficulty of removing ineffective teachers, a larger problem is how to keep promising teachers in classrooms and support them long enough so they become effective.

We see value in efforts to create a more educationally sound, more reliable system for evaluating teachers. However, the new teacher evaluation plan proposed by the Massachusetts Department of Elementary and Secondary Education fails to address the real challenges facing our public schools. It threatens to undermine our ability to recruit, retain and develop excellent teachers, especially in those schools and districts that need them most. (The plan will have a similar impact on principals.) We see five main problems with the proposal:

- **It will require districts to use MCAS results to judge educators**. MCAS tests were not designed for this purpose. Using them to evaluate teachers and principals will intensify teaching to the tests and further narrow and dumb down teaching and learning in our classrooms.

- **It will require districts to evaluate every teacher in every grade and subject with two "assessments" each academic year, forcing districts to make or purchase dozens of new tests.** This will be an enormous and irresponsible expense at a time of teacher layoffs and extensive cutbacks. Most of the tests are likely to be narrow and low quality. This additional testing requirement will create another layer of bureaucracy and red tape. It will eat up time, energy and resources desperately needed to focus on the main task at hand—helping teachers be as effective in the classroom as possible.

- **It incorporates a type of "value-added measurement" (VAM) that is unproven and likely to be counterproductive.** Independent experts in assessment have determined that VAM is so flawed and inaccurate it risks producing unfair and destructive outcomes for professional educators—the opposite of its stated intent.

- **This high-stakes use of MCAS results is likely to rupture essential relationships between teachers and students.** Its use will supplant the needs of children as individuals with the bureaucratic requirement for data and graphs.

- **Comprehensive, high-quality teacher evaluation systems already exist and are used in many schools and districts.** The problem is not the lack of good models, but the lack of resources, time, training, and focus needed to implement them.

Good teachers do not simply convey information. They identify the diverse needs of their students; they engage student interests and build students' confidence; they help develop team interaction and cooperation; they challenge their students and assist them in overcoming barriers. They listen to students and identify student issues and concerns. They enable students to think and use content knowledge.

The primary goal of teacher evaluation should be to provide assistance where needed, and to recognize talented teachers who can play a leading role among their peers. Where necessary, such a system should play a role in removing teachers who are not effective and do not improve despite the assistance.

We were encouraged by the values that informed the work of the *Massachusetts Task Force Report on the Evaluation of Teachers and Administrators*, including that:

- Student learning, growth and achievement extend beyond academic progress and include other developmental factors – social and emotional well-being, civic learning and engagement.

- Educator expertise is the foundation of educator effectiveness.

- Leadership, school climate and culture are essential elements for supporting the learning and growth of both students and adults.

However, the Department's proposal is fundamentally inconsistent with these values. It threatens to intensify the negative impact of a system already too focused on MCAS tests. Most importantly, it fails to create a system that incorporates comprehensive and multi-faceted measures of learning and well being among our schoolchildren without causing harmful collateral damage.

Brown University Professor Marie Myung-Ok Lee recently wrote a *New York Times* oped about how her high school English teacher, Ms. Leibfried, recognized and coaxed out the latent writer hiding within the shy, socially marginalized schoolgirl. Ms. Liebfried nurtured Lee's self-confidence and helped her to express herself. "If we want to understand how much teachers are worth," Lee wrote, "we should remember how much we were formed by our own schooldays. Good teaching helps make productive and fully realized adults – a result that won't show up in each semester's test scores and statistics."

In the remainder of this report, we provide details and references to support our call for the Department to withdraw its proposal or for the Board to reject it, in order to replace it with a fair, effective and educational beneficially method of educator evaluation.

Limitations of MCAS Tests as Measures of Student Learning

"One thing I never want to see happen is schools that are just teaching the test because then you're not learning about the world, you're not learning about different cultures, you're not learning about science, you're not learning about math. All you're learning about is how to fill out a little bubble on an exam and little tricks that you need to do in order to take a test and that's not going to make education interesting."
—President Obama, March 28, 2011

This report is focused on the problems and dangers of using student test scores to evaluate teachers. But before we consider the evidence on that particular use of standardized test results, we must point out that instruments like the MCAS have limited value even in their primary role of assessing students, which places their use for teacher evaluation on a weak foundation.

The Most Important Aspects of Learning Are Not Assessed by MCAS

To become productive citizens in a democracy, students need to develop rich and complex reasoning powers and skills to deal with the complexity of the real world in which they live, the ability to assess the dimensions of a problem and pose the key questions, skill and practice in accurately observing the natural world, and the ability to work together cooperatively to reach a goal.

Preparing for standardized exams does not build the skills needed for college and employment. On the contrary, a focus on learning out-of-context facts and skills to pass exit exams detracts from preparing students for the work required in college. A survey of professors and employers by Achieve, an organization that promotes standards and tests, found many high school graduates are weak in comprehending complex written materials, communicating orally, understanding complicated materials, doing research, and producing quality writing (Achieve, 2005). Regardless of whether or not high schools prepared students well before high-stakes testing, such testing has been widely adopted in the past two decades, is now the status quo, and has clearly not helped achieve these goals.

Other flaws of high-stakes standardized tests have been well documented by numerous researchers. Briefly, they include these:

- **The high-stakes MCAS measures a narrow range of skills and knowledge:** Even for the tested subjects, MCAS measures too little and in too narrow a way, so that measured gains on MCAS at best only partially indicate whether they have gained on a full range of knowledge and skills, and at worst are seriously misleading (AERA, 2000; Berliner, 2007).

- **The high-stakes MCAS promotes test preparation:** Schools across the Commonwealth vary greatly in the form and quantity of test-prep — sometimes in special classes, sometimes after school, sometimes replacing regular classroom sessions. This form of "teaching" is far from the goal of preparing students to observe care-

High Stakes Testing Weakens Science Education

As a professor of molecular biology at MIT, I can attest to the damaging influence of the high-stakes MCAS on science education. Pressure on teachers to have their students perform well on standardized tests reduces the classroom role of experimentation, the design and construction of projects, field trips, and related encounters with natural processes. By shifting emphasis from direct encounters with natural phenomena to test preparation, high-stakes exams have helped alienate students from science and technology and turned science education back to pre-World War II, rote-learning modes.

Though the science MCAS tests assess whether a student knows the names of the parts of a microscope, they do not assess whether the student can actually focus the microscope, observe the sample, and trust their own observations enough to interpret the image. These are the kinds of skills that our students and society need most. I see this in some of my recent undergraduate students, who often think their task is to determine the 'right' answer, rather than to tell me what they observe.

—Jonathan King, Prof. of Biology, MIT

fully, think critically, and interact cooperatively and productively. More important even than the time spent on direct MCAS prep is the distortion of regular classroom instruction to focus on the MCAS tests. A prime example is the inordinate attention paid and time spent training students to write a "five-paragraph essay," a poor substitute for exploring the full range of skills required to become a competent and persuasive writer.

- **All such tests have measurement error:** This means an individual's score may vary from day to day due to testing conditions (for example, a barking dog outside the classroom window) or the test-taker's mental or emotional state (Hembree, 1988). As a result, many individuals' scores are frequently wrong. Test scores of young children and scores on sub-sections of tests are much less reliable than test scores on adults or whole tests.

In summary, we do not doubt that standardized tests capture some of the learning that takes place in a classroom. However, for many students this is the least important aspect of the learning needed to be productive citizens in a modern society (Dee & Jacob, 2006). In lower income communities, students face difficulties beyond the ability of the teacher to control, and these conditions have a substantial impact on their test scores (Heffley, 2007). While good teachers and schools can make a big difference in a student's life, they cannot completely make up for the inequities in society. Thus, to judge teachers erroneously on this basis will often deprive students of talented teachers and keep ineffective teachers in the classroom.

References

Achieve (2005). Rising to the challenge: Are high school graduates prepared for college and work? http://www.achieve.org/files/pollreport_0.pdf

AERA (2000). AERA Position Statement on High-Stakes Testing in Pre-K – 12 Education. *American Educational Research Association*. http://www.aera.net/policyandprograms/?id=378

Berliner, D. & Nichols, S. (2007). *Collateral Damage: How high-stakes testing corrupts America's Schools*. Cambridge, MA: Harvard Education Press.

Dee, T.S. & Jacob, B.A. (2006, April). Do high school exit exams influence educational attainment or labor market performance? Social Science Research Network. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=900985.

Haney, W. (2000, August 19). The Myth of the Texas Miracle in Education, *Education Policy Analysis Archives*. epaa.asu.edu/epaa/v8n41/.

Heffley, E. (2007). What do CAPT scores really tell us? *The Connecticut Economy*, Summer: 14-16.

Hembree, R. (1988). Correlates, causes, effects, and treatment of test anxiety. *Review of Educational Research*, V58, N1, 1988.

Neill, M. (2005, August 4). School Beat: Notes on the recent NAEP test results. *Beyond Chron*. http://quartz.he.net/~beyondch/news/index.php?itemid=339.

Preparing for standardized exams does not build the skills needed for college and employment. On the contrary, a focus on learning out-of-context facts and skills to pass exit exams detracts from preparing students for the work required in college.

Too Many New Tests

The state's proposed teacher evaluation regulations will require districts to make or purchase dozens of new tests so that teachers in all subject areas and all grades, even kindergarten, can be evaluated. 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. Therefore, they will be forced 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 further dumb down curriculum and instruction, ensuring our students will be less prepared to be effective citizens, successful in college or at well-paying jobs, or 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 and look at representative samples of the ongoing work of each student. However, 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 based on 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.

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. Clearly, MCAS does not assess most of those attributes. There is no reason to believe the comprehensive and higher-order skills sought by colleges and employers will be measured by or embedded in new commercial or Department-made tests.

In short, the proposal is based 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, including performance tasks, 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.

In the best of circumstances, districts would work with their teachers to create and score high-quality performance assessments and look at representative samples of the ongoing work of each student. However, few districts will be able to do so, particularly with their now-diminished resources.

The Fatal Flaws of Value-Added Measurement

Data-based decision making is an indisputably important practice in public education. However, few things are more damaging to public confidence or unfair than a public policy that mandates the use of invalid or unreliable data to make important decisions, such as determining a teacher's or principal's evaluation, pay, or dismissal. Using untrustworthy data, even partially, to evaluate teachers is worse than using no data at all. Such an approach can lead to the wrong conclusions about teacher competence, as well as damage teaching, learning and school climate. For the sake of our public school students, we simply cannot afford to use a system that will punish or push out some of the best teachers while rewarding some of the least competent. However, that is the probable outcome of teacher evaluations based on value-added measurement.

Value-added measurement (VAM) is a relatively recent approach to evaluating teacher performance. It is intended to measure a teacher's unique contribution to his/her students' academic progress. In other words, the purpose is to provide a single number representing the impact of a teacher on student achievement. This practice may sound good in theory, but it is fraught with problems that have yet to be solved. Indeed, most of the recent and methodologically strongest research on this topic (e.g., Corcoran, 2010; Lockwood, McCaffrey, Hamilton, Stecher, Le & Martinez, 2007; Papay, 2011; Rothstein, 2011; Schochet & Chiang, 2010) indicates that current approaches to VAM are so badly flawed that they should have no role in teacher evaluation.

Value-Added Scores Are Unstable

One critical flaw is that VAM-based teacher performance scores are highly unstable. VAM scores, for example, have been found to fluctuate widely depending on the specific test items used to measure the subject areas of reading (Rothstein, 2010) and mathematics (Lockwood et al., 2007), as well as the time of the year at which the tests are administered (Papay, 2011). Although Kane (2010) found that VAMs based on different student achievement tests in the same subject area "tend to" be correlated with one another, the correlations are so weak that they render the value-added assessment unusable. As Rothstein (2011) pointed out, about 45% of the teachers in Kane's study (2010) scoring above average (i.e., at the 80th percentile) based on one measure of student academic achievement in a subject area would score below average using a different measure of student academic achievement in the same subject area. This suggests that VAM cannot untangle the effects of an individual teacher from the varied results that come from different achievement tests used to measure teacher effectiveness. This is an unacceptable flaw in any evaluation system.

Teachers Judged on Factors They Do Not Control

VAM also holds teachers accountable for factors that are beyond their control. In Massachusetts as well as most other states, there is no direct measurement of student growth from September through June, the period of time during which teachers can directly impact student learning. MCAS is annually administered to most students in March. This means that when we examine year-to-year growth rates, we are not only assessing what occurred during that school year but also what has occurred during the previous summer and during the previous spring when a student had another teacher. This is a significant problem for VAM because many students experience significant academic losses or gains during the summer months (Cooper et al., 1996). Children from affluent families often spend summers engaged in artistic pursuits, athletics, travel or educational activities, whereas children from financially disadvantaged homes often lack access to any kind of enrichment and may actually lose academic skills during the summer. Papay (2011) found that VAM-based performance scores vary widely depending on what time of year (i.e., fall or spring) the tests are given. One of the probable causes of the instability in VAM scores was the academic gains and losses that occurred during the summer months.

Statistical Challenge Makes VAM Hardly Better than a Coin Toss

The basic assumption of VAM--that it is possible to reliably and accurately identify a teacher's unique

contribution to student learning--is questionable. The challenge is all the more daunting given that teachers probably influence less than 10% of the variation in students' progress in academic achievement (Schochet & Chiang, 2010). The best approach to identifying a teacher's unique influence on students would be to randomly assign students to classrooms and schools.

However, in the real world, students are anything but randomly distributed in classrooms. They come from neighborhoods and attend schools with widely varying income and ethnic compositions. Some are enrolled in programs or classrooms that specialize in addressing various special needs or language learning requirements. Some attend schools that "track" their students (explicitly or implicitly) according to perceived academic "ability." Therefore, since teachers seldom teach a randomly selected group of students, researchers must try to statistically control for all the important factors (other than teacher performance) that might impact changes in student achievement scores, such as poverty, disability, and English proficiency. They attempt to create statistical estimates of the teacher's unique contribution to students' academic progress. However, these estimates again are full of flaws, and inevitably do an inadequate job of accounting for some critically important factors. (To use just one example of many: While a particular method might compare teachers who had a certain percentage of students on Individual Education Plans, it might make no distinction between students with severe cognitive or emotional issues vs. students with minor learning disabilities.) After reviewing this topic, Baker et al. (2010) asserted that "VAM results are based on factors other than teachers' actual effectiveness."

The flaws inherent in VAM led Rothstein (2011) to conclude that "Teacher evaluations based on observed state test outcomes are *only slightly better than coin tosses* [emphasis added] at identifying teachers whose students perform unusually well or badly on assessments of conceptual understanding." Using a different set of test data, Corcoran (2010) reached a similar conclusion and stated that the value-added approach is in its "infancy."

Most Teachers' Contributions Are Not Measured by VAM

Another troubling limitation of VAM is that less than 20% of teachers teach English or mathematics, the only subjects taken into consideration by VAM. What about the majority of teachers and other school staff

A Formula for Disaster for One Exceptional Teacher

Box I. A Value-Added Model for a Given Subject, Grade, and Year

$$Y_{1i} = \zeta + \lambda Y_{0i} + \beta' X_i + \sum_{l(\text{district})} \sum_{k(\text{school})} \sum_{j(\text{class})} \left\{ \alpha_{jkl}^{\text{class}^*} I_{Cijkl} + \alpha_{kl}^{\text{school}^*} I_{Sikl} + \alpha_{l}^{\text{district}^*} I_{Dil} \right\} + \varepsilon_i$$

Posttest
Pretest
Student Characteristics
True Total Classroom Effect
True Total School Effect
True Total District Effect
Stud. Error Term
Classroom Participation Indicator
School Participation Indicator
District Participation Indicator

This formula accompanied a powerful article by Michael Winerip in the March 6, 2011 New York Times about Stacey Isaacson, a Queens, NY, teacher praised by all as creative, dedicated and highly effective. "'Definitely one of a kind,' said Isabelle St. Clair, now a sophomore at Bard..." According to the VAM formula, however, Ms. Isaacson was deemed one of the city's worst teachers and was told, as a result, she would not get tenure.

who may influence children's well-being and success in school but have nothing to do with these subjects? Ask any parent and they can probably think of an extraordinary art, music, physical education teacher or school nurse who made the difference between a child's success or failure. It makes no sense to invest scarce resources, time and effort into developing and implementing a teacher evaluation tool that is not only inaccurate but cannot be used to evaluate most teachers.

The absurdity of using such a flawed and limited measurement system as public policy can be captured in the following analogy. Most of us trust opinion polls when we are informed that the results are accurate within three points. This means that if the opinion poll indicated that 56% of those surveyed were planning to vote for candidate X, then there is 95% probability that the true percentage is somewhere between 53% and 59% (56% plus or minus 3%). In using four consecutive years of student achievement data, Corcoran (2010) found that when the 95% confidence interval is applied to VAM of a teacher falling at the 56 percentile, the true percentile score could range from anywhere from 32 (much below average) to 80 (much above average). If we would not trust an opinion poll with this much error, why should we make decisions about teacher tenure, pay or dismissal based on such error-prone data?

False Allure of Scientific Accuracy

What is particularly dangerous about the efforts to impose value-added measures is that they falsely convey the precision of scientific accuracy, despite all the research evidence to the contrary. Indeed, the research is so overwhelmingly against the use of VAM that one of the world's foremost authorities on testing, the Board on Testing and Assessment of the National Research Council of the National Academies of Sciences, stated, "VAM estimates of teacher effectiveness should not be used to make operational decisions because such estimates are far too unstable to be considered fair or reliable." We could save considerable time, money and effort by tossing coins to make decisions about teachers. This would of course be ridiculously unfair, but at least the unfairness would be obvious to all.

References

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., Ravitch, D., Rothstein, R., Shavelson, R., Shepard, L. (2010). Problems with the Use of Student Test Scores to Evaluate Teachers. Economic Policy Institute.

Cooper, H. et al. (1996). The effects of summer vacation on achievement test scores: A narrative and meta-analytic review. *Review of Educational Research,* 66 (3), 227-268.

Corcoran (2010). Can Teachers be Evaluated by their Students' Test Scores? Should They Be? Annenberg Institute for School Reform at Brown University. www.annenberginstitute.org/pdf/valueaddedreport.pdf

Kane, T. et al. (2010). Learning About Teaching: Initial Findings from the Measures of Effective Teaching Project. Bill and Melinda Gates Foundation.

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

Neill, M. (2011). Testimony to Mass Board of Education against using student scores to judge teachers, FairTest. http://fairtest.org/testimony-mass-board-education-against-using-stude

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

Rothstein, J. (2010, February). "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement," *The Quarterly Journal of Economics*, MIT Press, vol. 125(1), 175-214.

Rothstein, J. 2011, January. Review of "Learning About Teaching." National Education Policy Center. http://nepc.colorado.edu/thinktank/review-learning-about-teaching.

Schochet, P. Z., and Chiang, H.S. (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.

The Social Impact of Teacher and Student Accountability Schemes

Excerpted from an essay by James Horn and Rachel Squires Bloom

[Editor's note: James Horn and Rachel Squires Bloom base their essay (Horn & Bloom, 2011) on evidence collected during a two-year study (Horn, 2003, 2007) of Louisiana educators at an urban K-5 school and on unpublished data from a Massachusetts elementary school. Based on evidence of ruptured relationships between educators and students observed in Louisiana and Massachusetts, Horn and Bloom write in this essay that imposition of teacher evaluation based on student scores would raise the stakes even further and could make that rupture permanent.]

The social impact of high stakes testing in poor urban schools has grown particularly acute over the past 10 years. Besides the shrinkage of school curriculums to fit boundaries defined by annual tests, the disappearance of recess and play, and the elimination of engaged minds-on and hands-on teaching and learning, our research in poorer schools has uncovered another tragic outcome to high stakes testing: the elimination of care as the ethos that has bound together teacher and student.

Ironically, that erosion has been due to teachers' efforts to emphasize the positive and to avoid the negative in the face of the threat and reality of failure to make Adequate Yearly Progress. For as teachers learned to focus on the promised improvements under NCLB's framework, rather than on the tangible failures of many and sometimes most of the children in their classes, the needs of children as fragile human beings became supplanted by an image born of data and graphs and achievement scales. Avoiding failure seemed to require the full embrace of the ideology of testing, and the advocacy for children that was the core of teaching began to be rooted out by the false promise and a failed effort to "leave no child behind."

There is temptation to focus on student groups whose scores could be boosted to whatever number or category is deemed worthy of enhancing a paycheck or avoiding a sacking. In some classrooms, students on the lowest and higher ends of ability and achievement may be neglected, as they would provide less "bang for the buck." One teacher spoke facetiously or cynically (it is hard to tell the difference these days) of how students entering her classroom may be labeled as "pay cut" or "bonus." This is harsh, but the reality is that a model that explicitly ties children's scores to monetary worth creates such an atmosphere. Even caring, effective and empathetic teachers would to some degree be aware of how individual students may influence their pay.

High-achieving students could be left behind as well if teachers believe that effort is better spent on students whose scores can be edged up more easily. And even though most teachers still take seriously, or want to take seriously, their responsibility for the academic and emotional progress for every student in their class, tying job security or pay raises to test scores promises to further damage one of the most important relationships with adults that children have outside of family.

As the principal, now retired from a Title I school in Louisiana, summed up what happened to her school after the high stakes hurricane hit in 2000, "caring went out the window." It was sucked away by the endless swirl of stress and the anxiety of testing and test prep. Left in the wake now remains a sense of helpless loss and corrosive aggression for both teachers and children, as they pick through the remains to see what can salvaged of their human dignity after this ten-year blow.

References

Horn, J., & Bloom, R. S. (2011, March 31). Cash incentives: Yet another way to destroy quality education. CommonDreams.org. http://www.commondreams.org/view/2011/03/31-5

Horn, J. (2007). The LEAP for accountability? Ideology and practice of testing in a Louisiana urban elementary school. In M. Brown. (Ed.), *Still not equal: Expanding educational opportunity in society* (pp. 111-143). New York: Peter Lang.

Horn, J. (2003). LEAP-ing toward accountability: Ideology, practice, and the voices of Louisiana educators. *The Qualitative Report,* 8: 2, 224-250. http://www.nova.edu/ssss/QR/QR8-2/horn.pdf

Authentic and Comprehensive Teacher Evaluation

"Student learning and growth are about more than numbers. Making strong connections with ALL the diverse learners who are my students, motivating them, making sure they really understand, raising their expectations of themselves, collaborating with their families – these are hard to measure, but they are essential to my success as a teacher. We need a common understanding that students' academic growth and progress is not a linear equation, and we need an evaluation strategy that honors this complexity."
–Teacher of the Year and Member, Massachusetts Task Force on the Evaluation of Teachers and Administrators

"So, how do you judge teachers? Test scores are a mechanical fix to a human problem. We need the qualified judgments of experienced professionals, not test scores."
--Diane Ravitch, in a speech at Boston College, Dec. 1, 2010

So what would an authentic teacher evaluation system look like? If relying on student test scores is not wise, how should principals and other evaluators assess teacher performance? We don't have to look far for the answer. Comprehensive, high-quality teacher evaluation systems have been developed and are being used in many schools and districts across the state and country.

The problem, according to both teachers and administrators, is not in the lack of good models, but in the lack of resources, time, training, and focus needed to execute those models. Many agree that if the evaluation systems were used properly, by trained and experienced educators, they would stimulate and improve the work of all teachers, and would help identify those few who were not effective, did not improve with assistance, and needed to be counseled out of the profession.

The Task Force on Evaluation of Teachers and Administrators (2011) drew from these systems in making some of its recommendations to the Massachusetts Board of Elementary and Secondary Education. We applaud the effort to develop a framework to guide and inspire individual school districts—working in collaboration with local teachers' unions—to construct evaluation systems appropriate for each community. However, while we agree with some of the Task Force's recommendations, we firmly reject the notion that a reliance on student MCAS scores is a valid teacher evaluation strategy and we ask the Board to promote a broader and deeper approach to judging teacher effectiveness.

Comprehensive evaluation systems address the problems inherent in pro-forma, on-the-fly, subjective evaluations by a single harried administrator. They do so without resorting to the use of student test scores. As Diane Ravitch notes, most good evaluation systems have at their heart the observations and analysis by trained and experienced educators (2010). While student test scores and teacher rankings seem to provide "hard data" from outside sources, in fact that data is fraught with inaccuracies, as described earlier in this report. The actual "hard data" consists of the deep, rich, hard-to-quantify but very real information that is gathered in the course of skillful classroom observations, examination of teaching artifacts, and thoughtful discussions with teachers being evaluated.

A full description of good evaluation systems is beyond the scope of this report, but a number of common elements have emerged over the years. Charlotte Danielson has played an important role. In the 1990s, while working at the Educational Testing Service, she developed a framework for comprehensive systems that has been adopted and adapted by many since then (Danielson, 2001). Variations on the elements below have been described by Danielson in "New Trends in Teacher Evaluation," Anthony Cody in "A Quality Teacher in Every Classroom: New Report Takes On Evaluation" (2010), retired Lincoln-Sudbury history teacher Bill Schechter in a survey of current practices in some successful suburban school systems, and the report of the Task Force itself. See the appendix for more detailed descriptions of effective evaluation systems.

Henry Adams did not say, "A teacher affects eternity; he can never tell where his influence stops" because he believed that high school students would forever remember their trigonometry. Adams understood that the influence of a teacher in a child's development is far-reaching. Yes, teachers must explain the Pythagorean theorem and the Hay-Pauncefort Treaty and give homework that helps to consolidate academic skills. But teachers must also inspire interest, help build character, stimulate curiosity, challenge assumptions, ask thought-provoking questions, model the pleasure and excitement of learning, exhibit and impart organized habits of mind, demonstrate critical thinking, ignite the imagination, deepen the virtues of responsibility and self-discipline, encourage compassion, foster citizenship, community, and respect for all in a democratic society. And, in all of this, they need to create classrooms where students can believe in themselves and find courage to try, even where failure is a possibility. This is why Adams believed a teacher "affects eternity."
—Bill Schechter, Lincoln-Sudbury High School history teacher, retired

Comprehensive evaluation systems begin with a set of explicit standards, understood and agreed to by all. Danielson broke teaching into four major categories, each with a number of elements, to be evaluated: 1) planning and preparation: knowledge of content and pedagogy, knowledge of students, design of coherent instruction, etc. 2) classroom environment: creating a culture of respect and rapport, a climate for learning, managing student behavior, organizing physical space, etc. 3) instruction: communicating with students, engaging students, using skillful questioning and discussion techniques, etc. and 4) professional responsibilities: reflecting on teaching, maintaining accurate records, communicating with families, participating in a professional community, etc.

She also created a scoring rubric—with the categories distinguished, proficient, basic, and unsatisfactory—by which to rate the skills of teachers. Multiple calibrations, rather than the traditional "thumbs up or down" approach, allow for a more nuanced assessment of the teacher's practice. The Task Force has suggested similar ratings for Massachusetts.

Classroom observations – both formal and informal, scheduled and non-scheduled – are at the heart of any evaluation system. The principal, another administrator, peer teachers, or sometimes outside observers visit the classroom and observe the teacher in action. This is done more frequently for new teachers or ones who need additional support, less frequently for more experienced and skilled teachers. (The frequency and terms of these visits are also usually spelled out in the teachers' union contracts.) The evaluator(s) then meets with the teacher and provides both written and verbal feedback.

In addition to classroom visits, the evaluator examines "artifacts" from the teacher's practice, which may include lesson plans, instructional materials, student work, videotapes of lessons, writing samples from the teacher's school-wide or district-wide professional activities, and evidence of work with parents. In some systems, such as the Connecticut BEST program for new teachers and the National Board program, specified materials are submitted as part of extensive portfolios.

Most effective systems call for teachers to be active, rather than passive, participants in the evaluation process. This approach often begins with the teacher reflecting on and describing her methods; conducting observed lessons; and interpreting events that occurred in the classroom. Teachers say that participating in their own evaluation challenges them to become more thoughtful, conscious, and skilled educators.

Successful evaluation systems pay close attention to training the evaluators. Evaluators should themselves be skilled and experienced educators with previous or current classroom experience (not business managers, for example). They should be trained in the standards, rubrics, and evaluation methods they will use so their judgments are accurate, consistent, and based on evidence. In some systems, the evaluators are primarily administrators; in others, negotiations with teachers unions have led to peer assistance review teams,

where teachers with extra experience and training take on evaluation as well.

The welfare and development of students is the ultimate goal of all efforts to improve schools, including a strong teacher evaluation system. While we object to the use of MCAS scores to evaluate teachers, we believe that student academic growth, stability, happiness, health, and achievement are goals that should permeate the entire evaluation process, and should inform each step. Teachers need to be able to manage classrooms so that students feel safe; to provide instruction so that students can learn content; to ask open-ended questions and stimulate discussion so that students feel included and engaged; to skillfully assess students' learning and provide meaningful feedback so that students are informed of their progress and challenged to improve; to provide understanding and support so that students can navigate often-difficult environments and stages of development.

In evaluating teachers, there are no shortcuts. We need to observe the way they discharge all of their responsibilities and "influence" the children they are educating. Their daily performances are held in a concert hall called the classroom.

Teaching is a sacred responsibility, so the rigorous evaluation of teachers is essential. No child should be consigned to a second-rate classroom by an incompetent teacher. Nor should a child be condemned to a second-rate education based on little more than test-prep drilling. Those who advocate a shallow evaluation for teachers or superficial test-based education for students would do well to reflect on these words of Edward Young so beautifully painted on the walls of the Library of Congress: "Too low they build who build beneath the stars."

References

Baker, E., et al. (2010, August). Problems with the Use of Student Test Scores to Evaluate Teachers. Economic Policy Institute. See pp. 20-21 for discussion of high quality teacher evaluation. http://epi.3cdn.net/b9667271ee6c154195_t9m6iij8k.pdf

Cody, A. (2010, June 1). A Quality Teacher in Every Classroom: New Report Takes on Evaluation, *Education Week*, http://blogs.edweek.org/teachers/living-in-dialogue/2010/06/a_quality_teacher_in_every_cla.html

Corcoran, S. (2010). "Can Teachers Be Evaluated by Their Students' Test Scores? Should They Be?" Annenberg Institute for School Reform. http://www.annenberginstitute.org/pdf/valueAddedReport.pdf

Danielson, C. (2001). "New Trends in Teacher Evaluation." http://charlottedanielson.com/articles.htm

Massachusetts Task Force on the Evaluation of Teachers and Administrators (2011, March). Building a Breakthrough Framework for Educator Evaluation in the Commonwealth.

Ravitch, D. (2010, December 1). Speech At Boston College: "So, how do you judge teachers? Test scores are a mechanical fix to a human problem. We need qualified judgments of experienced professionals, not test scores."

Ravitch, D. (2010). The Death and Life of the Great American School System. See especially Chapter 9: What Would Mrs. Ratliff Do? pp. 169-194.

Schechter, B. (2011, April). Evaluation Alternatives: Examination of Survey Results on Teacher Evaluation Methods in Fourteen Massachusetts Public School Systems.

Toch, T. & Rothman, R. (2008, January). Rush to Judgment: Teacher Evaluation in Public Education. Education Sector Reports. http://www.educationsector.org/usr_doc/RushToJudgment_ES_Jan08.pdf.

APPENDIX: Teacher Evaluation Alternatives In Massachusetts

by Bill Schechter

We have examined the evaluation plans and procedures of 14 public schools (or school systems) in Massachusetts that have long enjoyed the reputation of excellence both before and after the MCAS exams were introduced in the state. By certain measures, our state leads the nation in the academic achievement of its public school students. The schools referenced below are leaders in our state and can serve as models for schools everywhere.

Our examination demonstrates that great schools — with high student achievement and parent satisfaction — do not require teacher evaluations based on MCAS standardized test scores. Indeed, these schools have maintained the quality of their faculty by relying on careful in-classroom teacher observation.

The debate over the most effective models for teacher evaluation has been distorted by theory and ideology. Our study suggests that the debate would better be informed by the practical experience of excellent public schools over long periods of time.

The evaluation plans of these schools can be found online in the collective bargaining agreements negotiated between the school committees and union locals in their respective communities.

Because these plans are detailed and lengthy, we will only spotlight important commonalities as well as a few exemplary features. Readers can study the contracts in their entirety at the Department of Education website. The address is given below.

A. COMMONALITIES

1. Evaluation Schedule: All of the schools have detailed schedules as to a specific date for the beginning and the completion of the evaluation process, with dates as well for each component of the process: Planning meeting, goal-setting, observations, post-observation discussions, remediation plans, dismissal decisions, etc.

2. Frequency: For purposes of evaluation, all of these schools make a distinction between teachers with and without professional status. Professional status can be granted to a teacher only after three years of service. Teachers without professional status can be dismissed or nor re-hired without cause being given.

a. Without Professional Status:

All 14 contracts provided for the annual evaluation of teachers without professional status.
A few schools provided for two in-class evaluations, but most required three.
In all schools, teachers without professional status can be dismissed without any need to show cause.

b. With Professional Status

A few schools provided for in-classroom evaluations of teachers with professional once every 3 or 4 years. Most of these schools required formal evaluations every other year. Some required one or two in-classroom evaluations, and some required three. The frequency appears to have declined somewhat over the past contract cycle, reflecting the present generational shift in faculties, with its large influx of new teachers, with non-professional status, who require more intensive evaluation.

3. Length of In-Classroom Observations: A few stipulated 30 minutes at the elementary and junior high levels, but only one school system at the high school level. The others stipulated one full class period for each observation in high schools. No schools stipulated less than 30 minutes for any level.

4. Announced vs Unannounced Visits: About half of the contracts required announced visits; the others specifically provide for unannounced observations.

5. Off-Cycle Evaluation Years: All schools provide for an off-cycle evaluation process for teachers with professional status when they are not being directly observed. Generally, the off-cycle process involves an initial meeting to set goals, discuss professional development and school community involvement. Some schools provided opportunities for projects and portfolios, and for self- and/or peer evaluation. An end-year meeting is then required to assess progress made.

6. Rating Systems: Most of these schools use rating scales with three or four calibrations, ranging from "Satisfactory" to "Warning" for both professional status and non-professional status teachers.

7. Format of Evaluation: A few provide for a checklist option. Most use the narrative form.

8. Remediation Plans for Teachers With Warnings: All contracts provided for a plan to support teachers who receive performance warnings. The plans usually involved an increase in classroom evaluations, including off-cycle years, the listing of deficiencies, expectations, suggestions, process, and relevant dates. All contracts made clear a lack of progress could lead to dismissal, regardless of professional status.

B: SOME INTERESTING & NOTEWORTHY FEATURES

1. Criteria of Effective Instruction
- Acton-Boxoboro and Lincoln-Sudbury have a very detailed list.
- The Lincoln-Sudbury contract provides for all new teachers with a copy of Jon Saphier's book, *The Skillful Teacher.*

2. Evaluator Training
- All L-S evaluators share a common training in the Saphier method of evaluation.

3. Possible Materials That Can Be Submitted to Supplement Evaluation
- Again, Acton-Boxboro provides extensive options.

4. Pay Penalties
- In Brookline and several other systems, principals reserve the right to withhold pay increases from teachers who do not earn satisfactory ratings in the formal evaluation process.

5. Off-Cycle Year Evaluations for Teachers with Professional Status
- Newton's plan is particularly well-articulated.

6. Informal Evaluation Input
- L-S and several other schools note that "solicited" communications from parents, students, and colleagues can become part of the evaluation process, as well as "casual visits" to the classroom by administrators.
- Several schools require teachers to hand out student evaluation forms. Brookline is exploring with teachers how this information can become part of the formal evaluation process.

7. Relationship of Evaluation to Layoffs
- In the event of layoff due to a reduction in force, the Lincoln-Sudbury contract removes the preference normally accorded more senior teachers if they received unsatisfactory ratings.

8. Remediation Plans
- Several schools uses Peer Remediation teams as part of the remediation plan for teachers who receive unsatisfactory ratings.

9. Frequency

- Weston and several other schools "reserve the right" to make additional classroom observations beyond those stipulated by the contract and/or give teachers the right to request additional observations.

Links To 14 Massachusetts Teacher Contracts Referred To Above

Homepage: http://educatorcontracts.doemass.org/

ACTON-BOXBORO
http://educatorcontracts.doemass.org/view.aspx?recno=2

BELMONT
http://educatorcontracts.doemass.org/view.aspx?recno=23

BROOKLINE
http://educatorcontracts.doemass.org/view.aspx?recno=46

CONCORD-CARLISLE
http://educatorcontracts.doemass.org/view.aspx?recno=63

DOVER-SHERBORN
http://educatorcontracts.doemass.org/view.aspx?recno=74

LEXINGTON
http://educatorcontracts.doemass.org/view.aspx?recno=146

LINCOLN-SUDBURY
http://educatorcontracts.doemass.org/view.aspx?recno=148

LONGMEADOW
http://educatorcontracts.doemass.org/view.aspx?recno=150

NEWTON
http://educatorcontracts.doemass.org/view.aspx?recno=196

WAYLAND
http://educatorcontracts.doemass.org/view.aspx?recno=303

WELLESLEY
http://educatorcontracts.doemass.org/view.aspx?recno=305

WESTWOOD
http://educatorcontracts.doemass.org/view.aspx?recno=316

WESTON
http://educatorcontracts.doemass.org/view.aspx?recno=314

WINCHESTER
http://educatorcontracts.doemass.org/view.aspx?recno=325