FairTest

National Center for Fair & Open Testing

Teacher Evaluation Should Not Rest on Student Test Scores

The new federal Every Student Succeeds Act (ESSA) does not require states to have educator evaluation systems. If a state chooses to do so, it does not have to include student test scores. *To win federal Race to the Top grants or waivers from No Child Left Behind (NCLB), most states adopted teacher and principal evaluation systems based heavily on student test scores. Many educators have resisted these unproven policies. Researchers from Massachusetts and Chicago-area universities and more than 1,550 New York State principals signed statements against such practices. Chicago teachers struck over this issue, among others. Here's why these systems -- including "value added" (VAM) or "growth" measures -- are not effective or fair and hurt not strengthen teaching and learning.*

Basing teacher evaluations on inadequate standardized tests is a recipe for flawed evaluations. Value-added and growth measures are only as good as the exams on which they are based (American Statistical Association, 2014). They are simply a different way to use the same data. Unfortunately, standardized tests are narrow, limited indicators of student learning. They leave out a wide range of important knowledge and skills. Most states assess only the easier-to-measure parts of math and English curricula (Guisbond, *et al*., 2012; IES, 2009).

Test-based teacher evaluation methods too often measure the life circumstances of the students teachers have, not how well they teach. Researchers calculate teacher influence on student test scores ranges from as little as 1% to 14% (ASA, 2014). Out-of-school factors are the most important. As a result, test scores greatly depend on a student's class, race, disability status and knowledge of English. Some value-added measures claim to take account of students' backgrounds through statistical techniques. But the techniques do not adequately adjust for different populations or for the impact of school policies such as grouping and tracking students or the impact of race and class segregation on learning. So the measures remain inaccurate (Darling-Hammond, *et al*., 2012; Baker, 2013; ASA, 2014).

Basing teacher evaluations on VAM or growth harms educational quality. Determining educators' careers by their students' scores greatly intensifies incentives to narrow the curriculum and teach to the test (Guisbond, *et al*., 2012). More students lose access to untested subjects, such as history, science, art, music, second languages and physical education. Schools give less attention to teaching cooperation, communication, creativity, problem solving and other essential skills. Teachers also may avoid students whose scores are harder to raise (Mass. Working Group, 2012). Collins and Amrein-Beardsley's (2014) national overview found no evidence that using VAM data "works to improve instruction, much less increase subsequent levels of student achievement or growth." Another recent review of available evidence concluded that teacher evaluation has not produced positive results in learning outcomes or school improvement (Murphy, et al., 2013).

The huge number of new exams required to produce student scores used to judge all teachers gives new meaning to "testing overkill." Districts say they must add up to 1,500 hundred new tests to have data to evaluate the great majority of educators who teach courses other than math and reading (*Sun Sentinel*, 2014). Students and teachers were already experiencing test overload from federal, state and district testing mandates; now they face a veritable tsunami.

Because of unreliable and erratic results, many teachers are incorrectly labeled "effective" or "ineffective." On the surface, it makes sense to look at student gains, rather than students' one-time scores. However, VAM and growth measures are not accurate enough to use for important decisions. Newton, *et al.,* (2010) found that among teachers ranked in the top 20% of effectiveness in the first year, fewer than a third were in that top tier the next year. Another third moved all the way down to the bottom

P.O. Box 300204, Jamaica Plain, MA 02108
fairtest@fairtest.org 617-477-9792 http://fairtest.org

40%. High volatility in teacher VAM scores is the rule, not the exception (Adler, 2014). A RAND study found that using different math subtests resulted in large variations in teachers' ratings, suggesting the measure, not the teacher, caused the differences (Lockwood, *et al.*, 2007). In some states using these methods, the results are no more accurate than flipping a coin (Baker, 2012). Making changes despite such random variation can "can be detrimental to the goal of improving quality" (ASA, 2014).

It is difficult if not impossible to isolate the impact of a single individual on a student because teaching is a collaborative and developmental process. Teams of teachers, social workers, guidance counselors, librarians, school nurses and others work together. Classroom teachers also build on the efforts of previous teachers. If a student has a breakthrough in grade 5, it could be largely due to groundwork built in 3rd and 4th grade (Mass. Working Group, 2012). ASA also warns that using VAM could foster damaging competition and discourage positive collaboration among teachers (2014).

Use of VAM/growth models drives good teachers away from needy students or out of the profession. Excellent teachers can be judged "inadequate" by these tools; some leave the profession (Winerip, 2011a). Teachers working with the most needy students are put at risk because of their students' background characteristics (Burris, 2012; Mass. Working Group, 2012). Ironically, students who score highest on state tests also are likely to show little "growth," endangering their teachers (Pallas, 2012). It also appears that some districts are using teacher evaluation to remove older, higher-paid teachers, and particularly teachers of color, creating a whiter teaching force (Vaznis, 2013).

Many independent researchers conclude these methods are inadequate and will cause harm. VAM defenders claim the current teacher evaluation system is weak and must be changed. At a minimum, they say VAM will be better that what now exists. However, the Board on Testing and Assessment (BOTA) of the National Research Council concluded, "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" (BOTA, 2009). Bruce Baker (2011) summarized the research evidence: Value-added "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." (See also Haertel (2013) for similar conclusions from a research review.) But most states now put high, fixed weights on this data. As Baker (2012) says, the statistical models actually used by states "increasingly appear to be complete junk!" (emphasis in original).

Two high-profile studies often cited to support VAM fail to justify its use. A Gates Foundation study argued that teachers who scored high on VAM tend to do well on other measures (Kane, *et al*., 2010). Another study found that teachers whose students had high value-added scores also had students with better long-term outcomes such as higher incomes (Chetty, *et al.,* 2012). But several independent reviews found that neither study provided strong evidence that VAM's benefits outweigh the damage it can cause (Rothstein, 2012; Adler, 2014). Rothstein (2011) concluded the Gates report provided more reasons to *not* use VAM than to use it. Adler's (2013) analysis detailed five serious problems with the Chetty study and concluded that the only valid conclusion to be drawn is "the opposite of what's been reported." Both Gates and Chetty used data from teachers who did not face high-stakes consequences. Pressure to boost scores would likely corrupt the results, further undermining their arguments.

To evaluate teacher and principal quality and effectiveness, use multiple measures based on school and classroom evidence. To the limited extent that scarce resources should be spent on teacher evaluation (Murphy, *et al*., 2012), the fair and accurate way to determine an educator's quality is with an array of different measures (Mathis, 2012). States and districts should use techniques that do not rely on student test scores, such the Peer Assistance and Review Model (Darling-Hammond, *et al*., 2012; SRI, 2011). Evidence from districts such as Montgomery County, Maryland (n.d.) and Toledo, Ohio shows that peer review systems (which focus mainly on professional learning) can be fair and accepted by educators (Winerip, 2011b; SRI, 2011). They also can improve the quality of teaching and counsel out teachers who should be in a different profession.

The public policy solution is to end Race to the Top, NCLB waivers, and state programs that mandate the use of student test scores. Doing so will require a movement of educators, parents, students and other members of the community that can win these changes (Guisbond, 2014).

- Updated January 2016.

References

Adler, M. 2014. *Review of Measuring the Impacts of Teachers*. National Education Policy Center. April. http://nepc.colorado.edu/thinktank/review-measuring-impact-of-teachers

Adler, M. 2013. “Findings vs. Interpretation in ‘The Long-Term Impacts of Teachers’ by Chetty, *et al*.” *Education Policy Analysis Archives*, V. 21, N. 10, Feb. http://epaa.asu.edu/ojs/article/view/1264/1033

American Statistical Association (ASA). 2014. *ASA Statement on Using Value-Added Models for Educational Assessment.* April 8. http://www.amstat.org/policy/pdfs/ASA_VAM_Statement.pdf

Baker, B. 2013. “The Value Added & Growth Score Train Wreck is Here.” Oct 13. http://schoolfinance101.wordpress.com/2013/10/16/the-value-added-growth- ...

Baker, B. 2011. “Opinion: 7 reasons why teacher evaluations won't work,” NorthJersey.com. http://www.northjersey.com/news/education/evaluation_031311.html?page=al ...

Board on Testing and Assessment. 2009*. Letter Report to the U.S. Department of Education on the Race to the Top Fund*. The National Academies. http://www.nap.edu/openbook.php?record_id=12780&page=1

Burris, C. 2012. “New teacher evaluations start to hurt students.” The Answer Sheet. *The Washington Post*. September 30. http://www.washingtonpost.com/blogs/answer-sheet/post/new-teacher-evalua...

Chetty, R., Friedman, J.N. and Rockoff, J.E. 2011. *The Long-Term Impact of Teachers: Teacher Value-Added and Student Outcomes in Adulthood*. National Bureau of Economic Research. http://www.nber.org/papers/w17699

Collins, C. and Amrein-Beardsley, A. 2014. “Putting Growth and Value-Added Models on the Map: A National Overview.” *Teachers College Record.* file:///Users/lisa/Downloads/nat%20survey%20of%20VAM%20implementation.htm

Darling-Hammond, L., et al. 2012. “Evaluating Teacher Evaluation.” *Phi Delta Kappan*. March. http://www.kappanmagazine.org/content/93/6/8.short

Guisbond, L., Neill, M., and Schaeffer, B. 2012. *NCLB’s Lost Decade for Educational Progress: What Can We Learn from this Policy Failure?* FairTest. http://fairtest.org/NCLB-lost-decade-report-home

Institute of Education Sciences. 2009. Appendix A: State Testing Programs Under NCLB. http://ies.ed.gov/ncee/pubs/2009013/appendix_a.asp

Lockwood, J.R., McCaffrey, D.F., Hamilton, L.S., Stecher, B.M., Le, V. and Martinez, F. 2007. "The Sensitivity of Value-Added Teacher Effect Estimates to Different Mathematics Achievement Measures." *Journal of Educational Measurement*. Spring, V. 44, N. 1, pp.47-67.

Mathis, W. 2012. *Research-Based Options for Education Policy Making: Teacher Evaluation*. National Education Policy Center. http://nepc.colorado.edu/files/pb-options-1-teval_0.pdf

Montgomery County Public Schools. (N.d.) Professional Growth System. http://www.montgomeryschoolsmd.org/departments/development/teams/admin/a.. .

Murphy, J., Hallinger, P, and Heck, R.H. 2013. “Leading via Teacher Evaluation: The Case of the Missing Clothes?” *Educational Researcher*, V.42, N.6.

Newton, X., Darling-Hammond, L., Haertel, E., & Thomas, E. 2010. “Value-Added Modeling of Teacher Effectiveness: An exploration of stability across models and contexts.” *Educational Policy Analysis Archives*, 18 (23). http://epaa.asu.edu/ojs/article/view/810

Pallas, A. 2012. “Meet the ‘worst’ 8th grade math teacher in NYC.” The Answer Sheet, *The Washington Post.* May 16. http://www.washingtonpost.com/blogs/answer-sheet/post/meet-the-worst-8th...

Rothstein, J. 2011*. Review of Learning About Teaching: Initial Findings from the Measures of Effective Teaching Project*. Boulder, CO: National Education Policy Center. http://nepc.colorado.edu/thinktank/review-learning-about-teaching

Rothstein, J. 2012. “Let’s Not Rush Into Value-Added Evaluations.” Room for Debate, *The New York Times.* http://www.nytimes.com/roomfordebate/2012/01/16/can-a-few-years-data-rev...

SRI International. 2011. *The Search for Teacher Effectiveness: A Study of Exemplary Peer Review Programs.* http://policyweb.sri.com/cep/projects/displayProject.jsp?Nick=PARPeer

Sun Sentinel. 2014, Sept. 4. “Timeout needed on high-stakes tests.” http://articles.sun-sentinel.com/2014-09-04/news/fl-editorial-high-stakes-testing-fcat-0905-20140904_1_high-stakes-assessment-tests-florida-legislature

Winerip, M. 2011a. “Evaluating New York Teachers, Perhaps the Numbers Do Lie.” *The New York Times.* March 6. http://www.nytimes.com/2011/03/07/education/07winerip.html?pagewanted=all

Winerip, M. 2011b. “Helping Teachers Help Themselves.” *The New York Times*. June 6. http://www.nytimes.com/2011/06/06/education/06oneducation.html

Vaznis, J. 2013. “Union says teacher evaluation plan has race bias.” *The Boston Globe*. April 24. http://www.bostonglobe.com/metro/2013/04/23/boston-union-officials-black-and-hispanic-teachers-disproportionately-targeted-under-new-evaluation-system/LCghntHAh8zM2R8qPmYrzM/story.html

Working Group on Teacher Evaluation. 2012. *Flawed Massachusetts Teacher Evaluation Proposal Risks Further Damage to Teaching and Learning.* *http://fairtest.org/flawed-ma-teacher-evaluation-proposal-report-home*