Neither Fair Nor Accurate • Research-Based Reasons Why High-Stakes Tests Should Not Be Used to E

Winter 2010

Illustration: J.D. King

By Wayne Au

A pitched battle raged in my hometown of Seattle this fall. Superintendent Maria Goodloe-Johnson and the Seattle Public Schools district fought with the Seattle Education Association over their most recent teachers' union contract. At the heart of the dispute: Should teacher evaluations be based in part on student scores on standardized tests?

Seattle is not unique in this struggle, and it is clear that Superintendent Goodloe-Johnson takes her cue from what is happening nationally.

In August, for instance, the *Los Angeles Times* printed a massive study in which LA student test scores were used to rate individual teacher effectiveness. The study was based on a statistical model referred to as value-added measurement (VAM). As part of the story, the *Times* published the names of roughly 6,000 teachers and their VAM ratings (see sidebar, p. 37).

In October the New York City Department of Education followed suit, publicizing plans to release the VAM scores for nearly 12,000 public school teachers. U.S. Secretary of Education Arne Duncan lauded both the *Times* study and the NYC Department of Education plans, a stance consistent with Race to the Top guidelines and President Obama's support for using test scores to evaluate teachers and determine merit pay.

Current and former leaders of many major urban school districts, including Washington, D.C.'s Michelle Rhee and New Orleans' Paul Vallas, have sought to use tests to evaluate teachers. In fact, the use of high-stakes standardized tests to evaluate teacher performance à la VAM has become one of the cornerstones of current efforts to reshape public education along the lines of the free market.

On the surface, the logic of VAM and using student scores to evaluate teachers seems like common sense: The more effective a teacher, the better his or her students should do on standardized tests.

However, although research tells us that teacher quality has an effect on test scores, this does not mean that a specific teacher is responsible for how a specific student performs on a standardized test. Nor does it mean we can equate effective teaching (or actual learning) with higher test scores.

Given the current attacks on teachers, teachers' unions, and public education through the use of educational accountability schemes based wholly or partly on high-stakes standardized test scores and VAM, it is important that educators, students, and parents understand why, based on educational research, such tests should not be used to evaluate teachers.

Although there are many well-documented problems with using VAM to evaluate teachers, I've chosen to highlight six critical issues with VAM that are so problematic they alone should be enough to stop the use of high-stakes standardized tests for such evaluations. I hope these will be helpful as talking points for op-ed pieces, blogs, and discussions at school board meetings, PTA meetings, and in the bleachers at basketball games.

Statistical Error Rates

There is a statistical error rate of 35 percent when using one year's worth of test data to measure a teacher's effectiveness, and an error rate of 25 percent when using data from three years, researchers Peter Schoche and Hanley Chiang find in their 2010 report "Error Rates in Measuring Teacher and School Performance Based on Test Score Gains," released by the U.S. Department of Education's National Center for Education Statistics.

Bruce Baker, finance expert at Rutgers University, explains that using high-stakes test scores to evaluate teachers in this manner means there is a one-in-four chance that a teacher rated as "average" could be incorrectly rated as "below average" and face disciplinary measures. Because of these error rates, a teacher's performance evaluation may pivot on what amounts to a statistical roll of the dice.

Year-to-Year Test Score Instability

As Tim Sass, economics professor at Florida State University, points out in "The Stability of Value-Added Measures of Teacher Quality and Implications for Teacher Compensation Policy," test scores of students taught by the same teacher fluctuate wildly from year to year. In one study comparing two years of test scores across five urban districts, more than two-thirds of the bottom-ranked teachers one year had moved out of the bottom ranks the next year. Of this group, a full third went from the bottom 20 percent one year to the top 40 percent the next. Similarly, only one-third of the teachers 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.

If test scores were an accurate measurement of teacher effectiveness, "effective" teachers would rate high consistently from year to year because they are good teachers; and one would expect "ineffective" teachers to rate low in terms of test scores just as consistently. Instead, the year-to-year instability that Sass highlights shows that test scores have very little to do with the effectiveness of a single teacher and have more to do with the change of students from year to year (unless, of course, one believes that one-third of the highest ranked teachers in the first year of the study simply decided to teach poorly in the second).

Day-to-Day Score Instability

Fifty to 80 percent of any improvement or decline in a student's standardized test scores can be attributed to randomly occurring factors, according to Thomas Kane of Harvard University and Douglas Staiger of Dartmo College in their research report "Volatility in Test Scores."

This means that factors such as whether or not a child ate breakfast on test day, whether or not a child got in argument with parents or peers on the way to school, which other students happened to be in attendance wh the test, and the child's feelings about the test administrator account for at least half of any given student's st test score gains or losses. Some factors, such as a dog barking outside an open window, can affect an entire

Kane and Staiger's findings illustrate that using tests to evaluate teachers ignores the reality that a host of inc daily factors that are completely out of a teacher's control contribute to how a student performs on any given reward or punish a teacher based on such scores could literally mean rewarding or punishing a teacher bas well or poorly a student's morning went.

Nonrandom Student Assignments

The grouping of students—either within schools through formal and informal tracking or across schools throu socioeconomic class, and linguistic (ELL) segregation—greatly influences VAM test results, as 10 leading re in teacher quality and educational assessment highlight in their policy brief "Problems with the Use of Studen Scores to Evaluate Teachers," published by the Economic Policy Institute.

These researchers note that "teachers who have chosen to teach in schools serving more affluent students r appear to be more effective simply because they have students with more home and school supports for thei current learning, and not because they are better teachers."

Even when VAM models attempt to take into account a student's prior achievement or demographic characte models assume that all students will show test gains at an equal rate. This assumption, however, does not ne hold true for groups of students who historically have performed poorly on tests, for English language learne asked to become proficient in both a new language and a tested subject area, or for students with disabilities test-based rates of progress may be incomparable to any other student.

Nonrandom student assignment means that a teacher could be punished, dismissed, or lose tenure purely b course they teach or the school they teach in has a significant population of traditionally low-scoring students show variable or slower test score gains.

Imprecise Measurement

High-stakes, standardized tests are also unable to account for the complexities of learning (and, by extensio teaching). For instance, we know from the linguistic research of Steven Pinker and others that learning often a U-shape—that making mistakes is an integral part of the learning process. When children are tested, we ne know where on the U-shaped learning curve they might be, nor do we realize that their mistakes could be a v a natural learning process. When tests are used to evaluate teachers, it is possible that highly effective teach push students out of their cognitive comfort zones are penalized for provoking the deep learning that requires to make mistakes on the way to greater understanding.

Standardized tests are also too crude to account for the possibility of cognitive transfer of skills that students across different subjects. Using VAM, as the researchers in the above-mentioned Economic Policy Institute p explain, means that "the essay writing a student learns from his history teacher may be credited to his Englis even if the English teacher assigns no writing; the mathematics a student learns in her physics class may be her math teacher." In other words, we can never be certain which class and which teacher contributed to a g student's test performance in any given subject.

Out-of-School Factors

Out-of-school factors such as inadequate access to health care, food insecurity, and poverty-related stress, others, negatively impact the in-school achievement of students so profoundly that they severely limit what s teachers can do on their own, explains David Berliner, Regents Professor of Education at Arizona State Univ his report "Poverty and Potential."

Although it is clear from the research of Stanford University's Linda Darling-Hammond and others that teache absolutely pivotal role in student success, when we use high-stakes tests to evaluate teachers, we incorrectl that teachers have the ability to overcome any obstacle in students' lives to improve learning. Although good are critically necessary, they are not always sufficient.

To assume otherwise is to think that teachers (and schools) can somehow make up for the lack of housing, f safety, and living wage employment, among other factors, all on their own. The social safety net is the respor much broader socioeconomic network—not the sole responsibility of the teacher.

Politics, Not Reality

The reality of standardized tests is that they are too imprecise and inaccurate to measure the effectiveness o teachers. The sad thing is that testing experts, researchers, and psychometricians have known this for quite In 1999, for instance, the expert panel that made up the Committee on Appropriate Test Use of the National R Council cautioned that "an educational decision that will have a major impact on a test-taker should not be m or automatically on the basis of a single test score."

Yet two short years later, a bipartisan Congress and the presidential administration of George W. Bush pass Child Left Behind and its test-and-punish approach to school reform into law.

Although the Bush administration seemed to ignore educational research as a matter of policy (as illustrated NCLB's Reading First program and the advocacy of using phonics-only teaching methods that had little basis research), many hoped for something different with the election of President Obama.

Unfortunately, the Obama administration has sent a clear message: When it comes to high-stakes standardiz the research doesn't matter.

It hasn't mattered that, according to the above cited U.S. Department of Education report, "More than 90 perc variation in student gain scores is due to the variation in *student-level* factors that are not under control of the

It hasn't mattered that the National Research Council of the National Academy of Sciences has stated that "V estimates of teacher effectiveness should not be used to make operational decisions because such estimate too unstable to be considered fair or reliable."

It hasn't mattered that even the researchers who completed the *Los Angeles Times* study acknowledged tha were too unreliable to use as the sole measure of teacher performance (a point that the *Times* neglected to c articulate in their article).

Sadly, with Bush, now with Obama, politics and ideology trump educational research.

One would think that all of the policy makers, politicians, pundits, superintendents, talk show hosts, documen makers, business leaders, and philanthropic foundations so in love with the idea of using test score data to e teachers would be equally as passionate about accuracy. People's lives are at stake, and yet the "data" unde important decisions about teacher performance couldn't be shakier.

The shakiness of test-based VAM data illustrates that the current fight over teacher "accountability" isn't reall effectiveness. The more substantial public conversation we should be having about rising poverty, the racial resegregation of our schools, increasing unemployment, lack of health care, and the steady defunding of the sector—all factors that have an overwhelming impact on students' educational achievement—has been burie teachers and their unions have become convenient scapegoats for our social, educational, and economic wo

Yes, teachers' performance needs to be evaluated, but in a manner that is fair and accurate. Using high-stak standardized tests and VAM to make such evaluations is neither.

A former high school teacher, **Wayne Au** is a *Rethinking Schools* editor and assistant professor at the Unive Washington, Bothell Campus.

Value Added and Human Dignity

Rigoberto Ruelas attended Miramonte Elementary as a student and returned to work there for 14 years as a TA and then as a 5th-grade teacher. He almost never missed a day of work. But Sunday, Sept. 19

called in sick. His body was found a week later underneath a 100-foot-high bridge in the Angeles Natio Forest.

Suicide rarely has a single cause, but Ruelas had been distraught over the Aug. 14 publication of an ar in the *Los Angeles Times*: "Who's Teaching L.A.'s Kids?" Ruelas' brother Alejandro told KABC-TV that Aug. 14, "he kept saying that there's stress at work." According to parents and staff at Miramonte, the principal had been pressuring Ruelas intensely since then to improve his students' scores, despite clai Los Angeles Unified School District (LAUSD) officials that they sent a memo to principals stating that us the data to discipline teachers is against the contract.

To write the article, *Times* reporters Jason Felch, Jason Song, and Doug Smith filed a public informatio request with LAUSD to get test scores for the students of 6,000 3rd-, 4th-, and 5th-grade teachers. LAU complied with the request, although they had never before used or published student test data disaggregated by teacher. The *Times* had the data analyzed using VAM methodology (see main article put the results online. Hundreds of teachers were publicly labeled "most effective," "more effective," "average," "less effective," or "least effective" based solely on changes in student test scores. Ruelas w rated "less effective."

But, according to students, co-workers, and parents, nothing could have been further from the truth. Christian, a former student of Ruelas' now attending high school in LAUSD, explained: "For me, he was good teacher. My parents were shocked to hear this. A lot of parents had respect for him. He was alwa there, whether he was sick or not. He was always smiling. He was happy with the students, and friendly the parents. He taught well. I liked being in his class."

Mayra Vega had stayed in touch with Ruelas since leaving the school six years ago. "He just told me tw weeks ago that he was proud of me for applying to college. He would always help you, even if you were his student. He always made me feel good about myself, like when he told me to go ahead and wear m glasses at graduation. Thanks to him, I stopped confusing my b's and d's."

According to Mat Taylor, south area chair of the teachers' union, United Teachers Los Angeles (UTLA) teaches the toughest 5th graders. Those are the kids he wants, even though they may be the ones who the hardest to test."

Two weeks after "Who's Teaching L.A.'s Kids?" was published, the Los Angeles School Board voted to accept a proposal by Deputy Supt. John Deasy that VAM account for 30 percent of teachers' evaluatio Two weeks later, the California State Board of Education voted to create an online database to track teachers by student test scores. Other districts have enacted even more extreme rules: In Florida and Denver, VAM may account for up to 50 percent of evaluations.

UTLA is demanding that the *Times* take down the weblink that rates individual teachers by name. "Rigo family wants his death to be for something," said Taylor. The process of learning and human developm cannot be assigned a number value. We need to resist the attempts to commodify our children as stude and ourselves as teachers.

—Sarah Knopp