TEACHER magazine

Published: April 28, 2010

No Value Added: The Mismeasurement of Teaching Quality

By David B. Cohen

Back to Story

Last summer, the Obama administration announced the Race to the Top competition, and planted the seeds for some serious changes in American public school systems. This spring, those seeds are bearing a bitter fruit, with state legislatures around the country rushing to enact reforms that are aligned with the new federal priorities.

American teachers find themselves battered and weary—already in a reactionary and defensive mode after nearly a decade's worth of No Child Left Behind, we're now in a fight to preserve our profession. The worst part of it, as we see it, is that the dollars our states are chasing will do little to improve teaching or schools.

Why are many teachers so angry? We're tired of being the scapegoats for the failures of an entire system. An echo chamber of major media outlets and pandering politicians have inflated real problems into existential crises, using "shock doctrine" to pave the way for favored, simplistic solutions. Teachers are frustrated that the supposed solutions are reforms selected without a basis in research, touted without proven results, and legislated without teacher input.

Secretary of Education Arne Duncan strikes many of us as particularly tone deaf on some of the key issues of the day, including the use (and misuse) of standardized tests. "Data-driven" is the mantra in education today—regardless of the quality of the data or how contorted the use of it. And one of the cornerstones of Race to the Top is the mandate to link student achievement data to teachers, for purposes of teacher evaluation.

Woe to the teachers who dare point out the problems with this approach. We are branded defenders of the status quo, obstructionists operating out of self-interest rather than looking out for the children. Yet one the clearest admonishments against using test scores for teacher evaluation comes not from teachers, but from a **joint statement** by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education:

Tests valid for one use may be invalid for another. Each separate use of a high-stakes test, for individual certification, for school evaluation, for curricular improvement, for increasing student

motivation, or for other uses requires a separate evaluation of the strengths and limitations of both the testing program and the test itself.

Additional studies have found that if one attempts to rank teachers by using their students' test scores, even using value-added measurements, the rankings shift depending on the test used and the rankings fluctuate in unpredictable ways over time.

None of this surprises teachers, because we live every moment of our teaching lives in an environment far too complex for bubble tests and silver bullets. I have come to believe that our greatest obstacle in education reform is a kind of willful denial of that complicated reality.

Attempting to puncture that bubble of blissful ignorance, I recently wrote a series of blog posts at **InterACT**, a group blog produced by Accomplished California Teachers. In our efforts to advance good teaching, ACT will soon be publishing a policy brief that contains numerous recommendations for improving teacher evaluations, but at the moment, it is also urgent that we expose the flawed assumptions about the use of state tests for teacher evaluation.

A Lack of Understanding

My arguments on the InterACT blog have been built around a series of questions for policymakers and others who call for greater use of standardized tests: Do you understand my job, my students, or my school?

Regarding my job, state standards instruct me to develop students' skills in four broad areas: reading, writing, listening, and speaking. The state tests claim to measure some of the reading standards and some of the writing standards, so even if we take that claim at face value, tests could only possibly assess my teaching in portions of two out of four broad standards areas. But I do not accept the test's claims. Asking students multiple choice questions about reading is an assessment approach with only marginal value, and it is just plain preposterous to claim we can assess student writing without the students actually writing.

So are advocates really suggesting that my work can be properly evaluated based on an obscured glimpse at a fraction of my job? It appears so.

And do policymakers really understand my students? The policies they advocate require acceptance of a huge and totally unwarranted assumption that student performance on state tests reflects their understanding. There are struggling students who resent and distrust tests, and fill in random answers, and there are more capable students who simply tire out or lose focus after an hour, or two, or three, or four, or nine. There are highly capable students who work themselves to exhaustion on homework, but rush through state tests and settle for whatever score they'll get. And why not? The tests have no meaning or consequences for students, and the data devotees have no way of controlling for this all-important variable. Policymakers who ignore the problem or believe it doesn't matter are kidding themselves, and doing all of us a disservice.

I dedicated two blog posts to the question, "Do You Understand My School?" Here's my argument: Two of the most important tenets in the use of tests, and in conducting research, relate to sample size and randomization. If standardized tests have any value at all, it is partly due to their sample size. Give fifty students a test and you can't be sure that their performance will be representative of the larger group of students their age in their state. Give the same test to 50,000 students and you might be able to draw some conclusions. State tests are designed to work on that much larger scale.

In my group of fifty sophomores, each student represents two percent of the total. So slight

differences can add up to significant percentages; if five of my students share a cold or flu virus around the time of the test, ten percent of "my" scores are affected. To protect against test flaws, there are also many different versions of the tests, but no measures to ensure that those versions are evenly distributed among my small group of students.

Of course, referring to any group as "my" students overlooks the fact that they have many other teachers and classes, where students read texts and practice reading skills that relate to the test they'll take in "my" subject area. Some students also have academic support classes or tutoring. So-called value-added measures supposedly distinguish my influence from all these others, but only with enough data. If we can't even identify, let alone quantify all the influences of various teachers, tutors, and curricula, we cannot reliably determine how much data is enough, nor should we infer causation from any observed trends in the data.

Unreliable Measures

Another challenge for value-added measurement is that I don't teach a random fifty sophomores. I teach students whose schedules worked out so that they ended up in my class. Depending on the placement of honors or remedial classes elsewhere in the schedule, my sophomores might be clustered on either side of the median skill level for their grade. I might have more or less than the average number of students with learning differences, depending on a number of scheduling or administrative factors.

Small sample sizes that aren't randomized prevent us from assuming that my sophomore classes match those of my colleagues, or that my sophomore classes this year resemble my sophomore classes last year. And yet policymakers seemingly ignore these issues, forcing us to accept evaluations based on a deeply flawed process.

An even more egregiously faulty assumption is that schools are similar enough from year to year to permit value-added measurements of state tests. Next year's changes at my school will include the following: a new principal, new assistant principal, new district administrator(s) for curriculum and instruction, expanded block scheduling (from two days to four), new time allotted for student tutorials and studying, new time allotment for professional development, new technology resources, and major new construction at multiple parts of the campus, displacing almost a fifth of our classes. How could we possibly disentangle the effects of all these changes when we're examining test results?

Furthermore, to use those results for teacher evaluation, you must assume that these school changes do not affect teachers, or affect all teachers equally. Otherwise, you're running a chemistry experiment in variable conditions using a half-dozen unknown compounds and telling me the results are still valid. If you concede that next year is an exception and maybe we shouldn't rely on those scores, then I have to ask which year can be considered "normal" and serve as a benchmark? Do we have a reliable measure of institutional stability?

To those who argue that I make too much of exceptional circumstances, I respond that in education, exceptions—viewed collectively—are the rule. No student, class, teacher or school is without exceptional elements. Use the tests at the state level and it's possible to even those things out. Use state tests at the school level, ignoring the purpose for which they were designed, and you stumble over one exception after another.

I would like to assume that Secretary Duncan is familiar with this information. I would like to know why it doesn't seem to matter to him, or to any of the state-level policymakers and think tank analysts who insist on going down this path of bubble test mania. To give Duncan some credit,

I've heard him say that we need much better, more robust tests. But some of those tests already exist, so Duncan would be more credible on this issue if he would take steps to promote their use, rather than propel us towards fatally flawed teacher evaluations based on what he seems to admit are inferior measures.

David B. Cohen is a member of the Teacher Leaders Network, and a National Board-certified teacher in Palo Alto, California, where he teaches high school English and serves as an academic adviser. He helps to direct ***Accomplished California Teachers*** *and writes for their group blog, InterACT.*

WEB ONLY