# RIGHT TASK, WRONG TOOLS: THE FLAWED APPRAISAL OF AMERICA'S TEACHERS

W. James Popham

University of California, Los Angeles

Everyone wants the nation's students to be taught by good teachers. Indeed, no right-thinking American believes bad teachers should be staffing our schools. Instead, we need effective teachers to be applauded while ineffective teachers, if they can't be improved, are sent packing. Yet, despite these almost universally held preferences, many state education officials are currently installing teacher-evaluation programs destined to yield distorted estimates of a teacher's quality.

Why it is, you might ask, are so many state authorities now feverishly scurrying to adopt brand new teacher-evaluation programs when, for decades, almost no attention was given to polishing our states' teacher-evaluation systems? The answer is that our federal government has recently established a waiver program allowing states to reduce the number of their public schools identified as "failing" under provisions of the No Child Left Behind Act (NCLB). The fewer a state's schools that flop on NCLB, the better the state's schools look. And state education officials, as is true of officials everywhere, would prefer to be regarded as successful rather than unsuccessful. Accordingly, well over half of our states are now seeking federal waivers to soften NCLB's sting.

However, in order to qualify for one of these waivers, a state's officials must first agree to install a rigorous system to evaluate the instructional caliber of their state's teachers. Moreover, in that system the test scores of a teacher's students must play a prominent role in determining a teacher's quality. On the face of it, this waiver-requirement seems altogether reasonable. After all, why shouldn't we appraise teachers chiefly on the basis of how much their students have learned?

Nonetheless, lurking in this intuitively appealing strategy for appraising teachers is the following significant requirement: *Any teacher-evaluation system based on students'*

*test scores must employ tests that primarily measure what students have been taught by their teachers.* Regrettably, the tests being chosen for use in these waiver-spawned evaluation systems fail to satisfy this indispensible requirement. With few exceptions, the tests chosen for today's state-contemplated teacher-appraisal systems are simply the annual accountability tests (in reading, mathematics, and science) currently being used to satisfy states' NCLB requirements. Those tests are dysfunctional when employed to evaluate teachers. Let's see why.

The fundamental measurement mission of all educational tests is to help us arrive at valid, that is, accurate, inferences about the students we test. Teachers rely on education tests because they can't see what's going on inside their students' heads. To illustrate, teachers can't look at a student, even if they stare hard, and arrive at an accurate judgment about how well the student can read a train schedule, write an essay, or solve an algebraic equation. Those skills are *covert,* that is, are unable to be seen. Consequently, teachers must rely on a student's *overt* test responses to test items in order to arrive at inferences about a student's covert skills and knowledge. The more important the test-based decision at issue, the more evidence that's needed to increase the likelihood we'll arrive at truly accurate (valid) test-based inferences.

But here's where the architects of most currently proposed teacher-evaluation systems are making a serious mistake. Put simply, the tests they plan to employ in those systems were built for an entirely different purpose and, as such, are accompanied by no evidence that they are capable of distinguishing between well taught and badly taught students. In other words, today's NCLB accountability tests cannot tell us who is a good teacher and who isn't. To use those tests as important components of a teacher-evaluation is flat-out folly.

The accountability tests currently used in the U.S. were built according to a traditional test-development strategy that, for almost a full century, has been widely accepted by those who construct educational tests. The overriding measurement function of these traditional tests has always been to differentiate among test-takers performances so one test-taker's results can be contrasted with the performances of other test-takers. This, of course, is why parents usually receive a child's test results in the form of

*percentiles* so that the test performance of a student who scores at, say, the 92nd percentile has exceeded the performances of 92 percent of the students in that test's norm (comparison) group. Such comparative interpretations can prove useful to both teachers and parents, for there are many settings in which we need to know how students stack up against one another.

But in order for educational tests to provide the information from which students can be accurately compared, it is imperative there be enough "score-spread" so that, when item-by-item scores are totaled up, there will be sufficient differences among students' total-test scores so that fine-grained comparisons among test-takers will be possible. As part of a traditional test-developer's quest to create sufficient score-spread, some items on their tests end up having little to do with what students are taught in school.

To illustrate, one way enough score-spread occurs on an educational achievement test is if many of the test's items are linked to students' socioeconomic status (SES) or to students' inherited academic aptitudes (such as children's *innate* verbal or quantitative potentials). Because both SES and academic aptitudes are nicely distributed variables, items linked to either one of them will invariably lead to the sorts of score-spread so desperately needed by such tests.

For instance, when a test item is linked to SES, there will typically be certain content in the item that allows students from more affluent backgrounds to do better on the item than students from less affluent backgrounds. Similarly, if an item is linked to inherited academic aptitude, for example, in mathematics, we usually see that students who were born with stronger quantitative potentials will have an easier time answering that item than will those students who, because of limited quantitative aptitudes, will always find math problems to be vexing.

Yet, even though test items tainted by SES or inherited aptitudes will contribute to more readily comparable total-test scores, such scores will not represent what students have been taught in school. More accurately, scores on such traditional tests frequently represent what students have brought to school, not what they learned once they arrive.

It is not that traditional test developers malevolently set out to penalize less affluent students or students who were less fortunate in their gene-pool lotteries. Instead, SES-linked items and aptitude-linked items usually find their way onto educational achievement tests simply because such items do a great job in spreading out test-takers' total scores.

Because most states' NCLB accountability tests have been constructed using traditional test-development procedures, it is not surprising that these tests provide no evidence about their use in determining teachers' effectiveness. Based on the way such tests have been developed, however, it is unlikely that today's NCLB accountability tests are capable of distinguishing between successfully and unsuccessfully taught students.

But relying on essentially unproven tests to make decisions about the caliber of a state's teachers is patently indefensible. Remember, we are not talking about test results in the abstract. No, these results can be used to reward and/or fire teachers. When wonderful teachers are mistakenly discharged, while inept teachers are retained, who is it that really loses out? It is, of course, these teachers' *students* who, almost certainly, will be less well taught.

Can educational tests be devised that both measure students' attainment of worthwhile curricular outcomes while, at the same time, accurately differentiating between skilled and unskilled teachers? The answer is a definite yes, but only if we cease our knee-jerk adulation of traditional test-development procedures. Instead, we need to build into the test-development process a variety of along-the-way operations that will allow us, in the end, to have evidence that such tests have a demonstrated legitimacy in a state's teacher-evaluation system. Students' achievement should definitely be the dominant ingredient in any defensible teacher-evaluation program. We must measure it properly.

Submitted as a "Pearson Author's Contribution" to www.myeducationcommunity.com.

February 8, 2012