USING THE WRONG TOOLS TO APPRAISE EDUCATIONAL QUALITY

W. James Popham
University of California, Los Angeles

Everyone wants children to be well educated. Accordingly, for more than a half-century, U.S. federal and state policymakers have been carefully trying to evaluate the quality of our nation's schools. Regrettably, the bulk of those evaluative efforts have failed miserably. That's because—with few exceptions—we have been using inappropriate tests to measure how well our students have learned.

Whether the focus of an evaluation is a state's entire school system or a particular school's effectiveness, the chief evidence that's employed to determine educational quality are students' scores on achievement tests, such as the annual state accountability tests required by federal law. Clearly, the quality and quantity of what students have learned in a school should be a dominant determiner of that school's success. Yet, almost all the tests we have currently chosen to evaluate our schools are flat-out wrong for this mission.

To do an accurate job of evaluating the quality of instruction within schools, a test must be "instructionally sensitive." In other words, it must be able to distinguish between well taught and badly taught students. However, if you were to review the technical documentation accompanying the standardized achievement tests we now employ to evaluate our schools, you would find there is no evidence—none at all—that these tests are up to that important assignment. Chiefly, this is because they are not measuring what we assume they are measuring.

This mismatch has historical roots. During World War I, U.S. Army officials commissioned the American Psychological Association to construct a written exam for recruits—an intelligence test to help identify potential lieutenants to lead the troops in France. They wanted an aptitude test to identify "the best of the best." The resultant test was called the "Army Alpha," and it was administered to about 1,750,000 Army recruits. It presented them with a set of verbal and numerical multiple-choice tasks, then sorted test-takers by comparing their total test scores; those who scored the highest were sent to officer training programs.

The Army Alpha was a hands-down winner, and much of its success stemmed from its design. The difficulty levels of items were expertly varied in a way that spread out the resulting scores so that fine-grained distinctions could be made among test-takers.

After the war, large-scale testing was introduced to U.S. education in the form of standardized achievement tests intended to measure students' mathematical, language, and social studies knowledge. These tests were built using the same score-spreading procedures pioneered during the Army Alpha's development. One crucial element of those procedures was the inclusion of numerous items that many test-takers would answer differently.

One of the very best ways to ensure that a test item produces varied responses is by linking the options in multiple--choice items to students' levels of affluence. If some answer-choices contain content that's likely to be familiar to children whose families' wealth provides more diverse experiences, those affluent students will get more correct answers than will their less affluent classmates. The Army Alpha included these kinds of items, and such items continue to be featured in the tests used to evaluate schools today. Although such affluence-slanted tests may do a crackerjack job of spreading out scores so that students can be compared, those tests tend to measure where a school's students are socioeconomically. In short, the tests used to evaluate schools often assess what students bring to school, not what they are taught once they arrive.

The work of America's educational test development firms is guided by the *Standards for Educational and Psychological Testing,* a joint publication of the three U.S. professional associations most concerned with educational assessment: the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education. These standards carry great weight both in the field and in courtroom contests involving educational tests. The most recent edition of these standards, published in 2014, makes it unambiguously clear that when a test will be used for an important purpose—such as evaluating schools—there must be convincing evidence indicating that the test's score-based interpretations will be accurate, that is, valid. There must also be convincing evidence that the test has been designed to perform the job that we intend it to perform.

Because a test that's not instructionally sensitive can make weak schools look wonderful and stellar schools look shoddy, we dare not use instructionally insensitive tests to evaluate the quality of our schools. They are the wrong tools for the job. Better tests can be built for this crucial measurement mission. They must be.