Correcting a Harmful Misuse of Students' Test Scores

By Invited Contributor Listed Below on June 3, 2014 2:29 PM | 1 Comment

*Today's guest contributor is **W. James Popham**, Professor Emeritus at University of California Graduate School of Education and Information Studies.*

A widespread, yet unsound use of students' standardized achievement test results is currently being countenanced in the educational measurement community in the United States (U.S). I refer to the role that student test results play in evaluating the quality of schools and, more recently, the effectiveness of teachers, with tests that lack instructional sensitivity. Depending on the tests being used, this practice is--at best--wrongheaded and--at worst--unfair. The *evaluative* use of students' performance results from most of today's *instructionally insensitive* tests violates the single, most important precept of educational assessment, namely, validity.

To use students' test results to evaluate schools and teachers, it is necessary to validate the interpretations and uses of test scores (**Kane, 2013**). To satisfy this sacrosanct validity cannon of educational testing, *evidence* must be available that allows test-users to answer two questions: (1) How accurate are score-based inferences about test-takers apt to be? (2) How appropriate will be the real-world uses to which those inferences will be put, such as, evaluating the effectiveness of schools and teachers?

The higher the stakes associated with the use of an educational test's results, the greater should be the scrutiny given to both the accuracy of score-based interpretations and to the appropriate usage of the test's results. In this series, **Madhabi Chatterji's blog** also speaks to these validity issues.

Accordingly, if students' performances on educational tests are being employed to evaluate the success of *schools*, it becomes imperative that those tests are accompanied by evidence supporting the legitimacy of using test-results for accountability purposes. So, for example, if a state's tests are intended to measure students' mastery of official, state-approved curricular aims, then evidence should accompany those tests indicating that test-based inferences about students' mastery status regarding the state's curricular aims are, indeed, valid (that is, accurate). Moreover, if the test results are also employed to rank the state's schools according to their levels of instructional success, then evidence must also be on hand indicating that the tests can actually differentiate among schools according to their relative effectiveness. Similarly, if students' test scores are to be employed in evaluating the quality of *teachers*, then we need not only evidence showing that the tests measure what a particular teacher is supposed to be teaching, but evidence is also needed indicating that results from the specific test being used can distinguish between well taught and poorly taught students.

During the past decade or so, attempts have been made to determine if educational tests used for high stakes accountability purposes measure the right curricular aims--typically by using so-called "alignment" studies in which systematic judgments are made about the degree to which a test's items address the knowledge and skills that students are supposed to learn. Yet, essentially no serious efforts have been made to indicate whether the *tests* being used to evaluate schools or teachers are actually up to that job. And this is a huge omission.

What Are Instructionally Sensitive Tests?

Instructional sensitivity refers to a test's ability to provide results that allow us tell how well test-takers were taught. Although minor definitional differences exist, most educational measurement specialists who deal with this topic accept a definition along the following lines:

Instructional sensitivity is the degree to which students' performances on a test accurately reflect the quality of instruction specifically provided to promote students' mastery of what is being assessed (**Popham, 2013, p. 64**).

This conception of instructional sensitivity has an unabashed focus on the *quality of instruction.* Thus, if the inference to be made from a test's results centers on the effectiveness of instruction provided to students by an individual teacher or by a school's entire faculty, the validity of those inferences about instructional quality from an instructionally *insensitive* test would clearly be suspect.

Because of recent federal incentives, almost every state in the U.S. has adopted new teacher evaluation programs in which students' test scores must play a prominent role. Regrettably, almost all of the tests currently being employed in the U.S. to evaluate school quality or

teacher quality have been created according to traditional testing practices of providing test scores that sort and rank test-takers best.

Tests that Rank Are Not Built To Be Instructionally Sensitive

That's right, for almost a full century, creators of America's standardized tests have been preoccupied with constructing tests that permit *comparative score interpretations* among test-takers. There is an educational need for such tests, particularly for admitting candidates into fixed-quota settings when there are more applicants than openings. But the need for comparative score interpretations disappears when we use a test's results to evaluate the quality of instruction given by teachers or schools.

As it turns out, many of the test-development procedures that are most effective in creating traditional, comparatively oriented tests are likely to *diminish* a test's instructionally sensitivity. For example, suppose a state's teachers have, because of a strong emphasis from state authorities, done a crackerjack instructional job during the last few years in promoting students' mastery of, say, Skill X. Well, during the process of designing a test for making comparisons among student performances, it is quite likely that items on the test measuring the well-taught Skill X will be deleted from the test. This is because too many students will be scoring well on Skill-X items. As a consequence, those items do not contribute to spreading out students' total-test scores--a necessary property if the test is going to do its comparative-interpretation and ranking job well.

Consequences of Using The Wrong Tests For Teacher and School Evaluations

It is bad enough when traditional, instructionally *insensitive* tests are employed to evaluate the quality of our nation's schools. The insensitivity of those tests surely leads to unsound decisions about schooling, and many students get a far less effective education than they should. That's surely reprehensible.

But now, because of federal incentives, most of our states have installed teacher evaluation programs in which students' test scores play a major role. If the tests being used are instructionally insensitive, then the evaluative judgments made about many teachers will be flat-out wrong. Effective teachers, misjudged, will be forced out of the profession. Ineffective teachers, misjudged, will be regarded as acceptable and, therefore, will continue to supply less than sterling instruction to their students.

And all these adverse consequences flow from a patently fixable flaw. When we evaluate schools or teachers with students' scores on tests whose suitability for those tasks has not been demonstrated, we not only violate validity fundamentals, we also violate fundamentals of fairness. In November , 2013 the First International Conference on Instructional Sensitivity was held at Lawrence, Kansas , directed by Neal Kingston and colleagues, where several methods of determining a test's instructional sensitivity were described (see **Neal Kingston's blog**).

W. James Popham
UCLA Graduate School of Education and Information Studies

Categories: Validity High-stakes testing School accountability International assessments Test use Formative assessment

Tags: High-stakes testing International assessments School accountability Test use Validity