

Accountability Tests' Instructional Insensitivity: The Time Bomb Ticketh

By W. James Popham

Would you ever want your temperature to be taken with a thermometer that was unaffected by heat? Of course not; that would be dumb. Or would you ever want to weigh yourself with bathroom scales that weren't influenced by the weight of the person using them? Of course not again; that would be equally dumb. But today's educators are allowing their instructional success to be judged by students' scores on accountability tests that are essentially incapable of distinguishing between effective and ineffective instruction. Talk about dumb.

What's worse is that we are now racing toward the 2014 deadline of the federal No Child Left Behind Act, the point at which all students are supposed to have attained test-based "proficiency." But the 2002-2014 schedules that most states devised when establishing their goals for annual required numbers of proficient students will soon demand some staggering increases in how many students must earn proficient scores on state NCLB tests each year. These balloon-payment improvement schedules were, in most instances, adopted as a way of deferring the pain stemming from having too many state schools and districts flop in reaching their goals for adequate yearly progress, or AYP.

Such cunningly crafted, soft-to-start improvement schedules will lead in a very few years to altogether unrealistic requirements for improved test scores. Without such improvements, huge numbers of U.S. schools and districts will be seen as AYP failures. If the American public is skeptical now about the quality of public schools, how do you think citizens will react when, in the next several years, test-based AYP failure becomes the rule rather than the exception? Can you hear the ticking of this nontrivial time bomb?

How could American educators let themselves get into a situation in which the tests being used to evaluate their instruction are unable to distinguish between effective and ineffective teaching? The answer, though simple, is nonetheless disquieting. Most American educators simply don't know that their state's NCLB tests are instructionally insensitive. Educators, and the public in general, assume that because

such tests are “achievement tests,” they accurately measure how much students have learned in schools. That’s just not true.

Two types of accountability tests are currently being used to satisfy the No Child Left Behind law’s assessment requirements. About half of the nation’s NCLB tests consist of traditional, off-the-shelf, standardized achievement tests, usually supplemented by a sprinkling of new items, so that the slightly expanded tests will supposedly be better aligned with a particular state’s content standards. Other NCLB tests are made-from-scratch, customized standards-based accountability tests, built specifically for a given state. Let’s see, briefly, why both these types of tests are instructionally insensitive.

Traditional standardized achievement tests, such as the Stanford Achievement Test-10th Edition, are intended to provide comparative information about test-takers. So the performance of a student who scores at, for instance, the 96th percentile can be contrasted with that of students who score at lower percentiles. To accomplish this comparative-measurement mission, these tests must produce a substantial degree of “score spread,” so there are ample numbers of high scores, middle scores, and low scores. Most items on such tests are of middle-difficulty levels because such items, statistically, maximize score spread.

Over the years, however, many of these middle-difficulty items turn out to be closely linked to students’ socioeconomic status. More-affluent kids tend to answer these socioeconomically linked items correctly, while less-affluent kids tend to miss them. This occurs because socioeconomic status, or SES, is a nicely distributed variable, and one that doesn’t change rapidly; SES-linked items help generate the score spread required by traditional standardized achievement tests. When such tests are used as accountability assessments, however, they tend to measure the socioeconomic composition of a school’s student body, rather than the effectiveness with which those students have been taught. The more SES-linked items there are on a traditional standardized achievement test, the more instructionally insensitive that test is bound to be.

How could American educators let themselves get into a situation in which the tests being used to evaluate their instruction are unable to distinguish between effective and ineffective teaching?

The other type of NCLB accountability test used in the United States is usually described as a “standards-based test,” because such tests are deliberately built to assess students’ mastery of a given state’s content standards, that is, its curricular aims. In all but a few states, though, the number of content standards to be assessed is so large that there is no way to accurately assess—via an annual accountability test—students’ mastery of this immense array of skills and knowledge. Instead, each year’s accountability test must sample from the profusion of the state’s curricular aims. Such a sampling-based approach to annual assessment means that teachers end up guessing about which curricular aims will be assessed each year. And, given the huge numbers of potentially assessable curricular targets, most teachers guess wrong.

After a few years of incorrect guessing, many teachers simply give up on trying to mesh their teaching with what’s to be assessed on each year’s accountability tests. And when this happens, it turns out that the major determinant of how well a school’s students perform on accountability tests is the very same factor that governed students’ performances on traditional standardized achievement tests: socioeconomic status. Thus, even on customized standards-based tests, a school’s scores are influenced less by what students are taught than by what the students brought to that school. Most standards-based accountability tests are every bit as instructionally insensitive as traditional standardized achievement tests.

The instructional insensitivity of accountability tests does not represent an insuperable problem, however. Remember when, several decades ago, we began to recognize that there was considerable test bias in our high-stakes educational assessments? Once this difficulty had been identified, it was attacked with both empirical and judgmental bias-detection procedures. As a consequence, today’s educational tests are markedly less biased than were their predecessors. Once the test-bias problem had been identified, we set out to fix it—and in less than a decade, we did.

That's precisely what we need to do now. Using a mildly technical definition, a test's instructional sensitivity represents the degree to which students' performances on that test accurately reflect the quality of instruction specifically provided to promote students' mastery of what is being assessed. We need to discover how to build accountability tests that will be instructionally sensitive and, therefore, can provide valid inferences about effective and ineffective instruction. It may take several years to get the required procedures in place, but we need to get started right now.

In the short term, though, we must make citizens, and especially educational policymakers, understand that almost all of today's accountability tests yield an invalid picture of how well students are being taught. Accountability systems based on the use of such instructionally insensitive tests are flat-out senseless. We need accountability tests capable of distinguishing between students who have been properly taught and those who have not. Until such tests are at hand, we might as well relabel our accountability systems as what they are—elaborate and costly socioeconomic-status identifiers.