

INSTRUCTIONALLY SUPPORTIVE HIGH-STAKES TESTS

W. James Popham

University of California, Los Angeles

American schools are almost awash these days under an ever-increasing array of high-stakes tests whose results have important consequences for students or for the teachers who taught those students.

The motives underlying the evaluative use of high-stakes tests have, for the most part, been commendable. The educational policymakers who installed such tests were typically trying to create test-based accountability systems that could, over time, reform a state's schools. Unfortunately, however, the very tests that were thought to be the cornerstones of a sensible school-reform strategy have, in many instances, led to a serious decline in educational quality.

In many parts of the U.S., we now see children being curricularly short-changed because their teachers fail to cover any content not assessed by their state's high-stakes tests. In those same communities, we frequently find children who are being drilled almost relentlessly on the topics covered in the state's high-stakes tests. Such drudgery-laden practice sessions are certain to diminish the joy that most students might derive from learning. And, distressingly, we increasingly hear reports regarding educators who, beleaguered by external score-boosting pressures, have engaged in

improper test-preparation activities or dishonest test-administration practices. Although understandable, such unethical behavior by teachers sends a message to our students that society simply doesn't want sent.

A Traditional Measurement Mission

These adverse consequences of test-based accountability systems do *not* arise simply because tests are being used as measures of educational quality. Rather, these negative effects are a direct result of our using *the wrong kinds of tests* in today's educational accountability systems. Almost without exception, the kinds of high-stakes tests now employed in the U.S. are either nationally standardized achievement tests, such as the *Iowa Tests of Basic Skills* or the *Stanford Achievement Tests*, or state-specific standardized achievement tests constructed in the same way as were those national tests. The chief measurement mission of all these traditionally constructed achievement tests is to provide relative comparisons among test-takers so that we can discover Chris scores at the 87th percentile in mathematics, but only at the 39th percentile in reading. Such comparative information is useful to both teachers and parents, for it allows them to isolate, at least in general terms, a child's relative strengths and weaknesses.

But in order for traditional standardized achievement tests to provide accurate comparative interpretations, the creators of those tests are often obliged to employ test

items that are, in a meaningful way, antithetical to the evaluation of instructional quality—whether at the district, school, or teacher levels.

In order for traditionally constructed achievement tests to accomplish their primary function of permitting accurate comparisons among students, those tests must routinely yield sufficient differences in students' scores. Putting it another way, the tests must produce adequate *score-spread*. However, for practical reasons, an achievement test, for instance, an achievement test in mathematics, must be administered to students in only about an hour or so. This time-limit forces those who construct traditional standardized achievement tests to employ items on their tests that will most efficiently contribute to score-spread. Among such score-spreading items are those linked to students' socioeconomic status (SES) and those linked to students' inherited academic aptitudes (such as students' inborn verbal, quantitative, and spatial capacities). Yet, those *SES-linked* and *inheritance-linked* items on standardized achievement tests actually measure what students bring to school, not what they learn there.

Thus, we see that the creators of traditional standardized achievement tests, in order to accomplish the historic *comparative* mission of those tests, must often employ items that are patently inappropriate for evaluating instructional quality. The effectiveness of schools ought to be judged chiefly on the basis of what students have learned in school. Test items that measure what students *bring to school*, of course, can't really tell us how well students have been taught.

Are traditional standardized achievement tests useful? Of course they are. As indicated earlier, both parents and teachers can profit from knowing children's relative strengths in different subject areas. But are traditional standardized achievement tests useful for judging educational quality? No, they are not. Indeed, it is the inherent instructional insensitivity of traditional standardized achievement tests that has led directly to many of the test-triggered educational problems we now see around us. More specifically, we encounter such problems in settings where traditionally constructed achievement tests are functioning as high-stakes assessments. For example, if a state's schools are ranked according to their students' scores on a standardized achievement test, it is a certainty that teachers will perceive the need to boost their students' test scores. That's because, after all, those test scores appear to represent a school-staff's success—or lack of it.

But there's also a related, yet serious shortcoming of traditionally constructed standardized achievement test. In such tests—tests designed to produce the kind of score-spread that permits precise comparisons among test-takers—it is not really necessary to describe, in very much detail, just what is being measured by the test. More accurately, it is not really necessary to depict what skills and knowledge are being measured *at a level of descriptive clarity sufficient for teachers' instructional planning*. Although the developers of traditional standardized tests routinely attempt to have their tests coincide with educators' curricular preferences, and although the tests are accompanied by *general* descriptions of the sorts of skills and knowledge being assessed, the resultant descriptive information is woefully inadequate. It is woefully

inadequate, that is, when teachers are trying to design lessons intended to raise their students' test performances due to students' improved mastery of what the test's items *represent*.

And this lack of descriptive clarity is why, in many instances, teachers who are being pressured to raise test scores will sometimes conclude that their only realistic option is to organize their instructional plans directly around the actual items making up a high-stakes standardized achievement test. Unfortunately, this sort of item-focused instruction may raise students' scores on a particular standardized test, but there will often be no concomitant increase in students' mastery of the knowledge and skills that the test was supposed to assess. Test scores may go up, but students' mastery of what was tested will not.

Even the way we have chosen to label our standardized tests contributes to the confusion regarding these tests' role in the evaluation of schooling. "Achievement," to most people, refers to what an individual has accomplished, as a consequence of that person's expending some sort of meaningful effort. Naturally, then, an educational "achievement" test would seem to refer to what students have, through their efforts, *learned* in school. But, to the extent that a traditionally constructed achievement test actually measures what students are bringing to school, because of their SES and their inherited aptitudes, then it is apparent that an achievement test is really not a measure of students' *learning*. "Achievement" is really the wrong descriptor for traditional tests of

this sort because many items in these tests fail to measure what students have *achieved in school*.

Test Results Required

The use of traditional standardized achievement tests in educational accountability programs, as you can see, is both misleading and educationally harmful. Nevertheless, there is a definite need for tests that can determine how well our students have been taught. Moreover, those tests must be *standardized* in the sense that they are administered and scored in a consistent, predetermined manner. Those proponents of rigorous accountability programs want to appraise the educational quality of our schools with data collected by using the same kind of measuring stick that makes sense. And, although the quality of any educational activity can be determined by using a variety of relevant evidence, one significant criterion of instructional quality is surely *the amount and quality of students' learning*. Tests are required to determine how much students have learned.

But what if standardized achievement tests were available that did what their name implies? What if standardized achievement tests could be used that measured what students had actually learned in school? And what if those standardized achievement tests were deliberately constructed to improve the quality of teachers' instructional decision-making? In short, what if *instructionally supportive achievement tests* could be installed that would (1) yield the kind of credible information about school

quality needed for educational accountability programs and (2) provide instructional support to the teachers responsible for promoting students' mastery of what was being tested?

In mid 2001, an independent group of 10 assessment/instruction specialists was convened by the American Association of School Administrators, the National Association of Elementary School Principals, the National Association of Secondary School Principals, the National Education Association, and the National Middle School Association. That group, the Commission on Instructionally Supportive Assessment, issued two reports in late-October 2001. One Commission report identified nine requirements that an instructionally supportive accountability test must satisfy. The second report supplied illustrations of how the Commission's nine requirements could be satisfied.

In the remainder of this analysis, I will be addressing three of the nine Commission requirements I regard as particularly pivotal when creating high-stakes tests that are instructionally supportive. I encourage readers to consult both of the reports of the Commission on Instructionally Supportive Assessment.*

* The Commission on Instructionally Supportive Assessment. (1) *Building Tests That Support Instruction and Accountability: A Guide for Policymakers*, (2) *Illustrative Language for an RFP to Build Tests That Support Instruction and Accountability*. Washington, DC: Author, 2001. Both reports are available online at: www.aasa.org, www.naesp.org, www.principals.org, www.nea.org, and www.nmsa.org.

Instructionally Supportive Standards-Based Tests

In order for an achievement test to satisfy the twin functions (1) of supplying accountability evidence and (2) supporting teachers' instructional decisions, the test must be built in a particular way. I'll now describe three essential elements in the test-building process that would lead to the creation of an instructionally supportive test. And, because almost all of today's achievement tests are intended to assess students' mastery of a set of *content standards*, that is, the knowledge and skills we want students to learn, I will refer to this new genre of high-stakes assessments as *instructionally supportive standards-based tests*. Let's consider the first requirement of such tests.

Standard-by-standard reporting. If a state's curricular aims are genuinely important, and the state's teachers are striving to have their students achieve those curricular aims, then the state's teachers need to know *which* aims were achieved and *which* aims weren't. If a state's teachers aren't being successful in getting their students to master a particular content standard, then the state's achievement tests should be supplying evidence to indicate the state's students aren't mastering that content standard. And if a specific teacher's students aren't mastering particular content standards, then that teacher clearly needs to know this so the teacher can make instructional changes. Conversely, if students *are* successfully mastering a given content standard, then teachers need to find this out so they'll know that their instruction is working. What this boils down to is quite simple. If teachers do not receive test-

based feedback *about individual students' mastery* of state-approved content standards, *on a per-standard basis*, then the instructional yield from those tests is destined to be negligible.

In most states, teachers receive only rather general feedback regarding students' overall mastery of a large collection of heterogeneous content standards. In those states, teachers are forced to *guess* about the appropriateness of the instruction they have provided related to different content standards. Yet, if standards-based assessment is supposed to underlie genuine educational reform, how can it do so if standards-based assessments fail to inform teachers about their success in promoting students' mastery of particular content standards?

It is only assessment evidence provided on a per-standard basis that can really illuminate teachers' instructional decision-making. Current claims that off-the-shelf tests are sufficiently "aligned" with a state's content standards tend to mask the reality that many of the state's content standards are not being measured at all or, at least, are not being measured in such a way that per-standard evidence of a teacher's instructional success can be provided. Moreover, even in those states that have attempted to build customized achievement tests to measure their state's content standards, because of the typical need to assess too many content standards during relatively brief assessment sessions, there is usually a failure to provide instructionally helpful per-standard evidence of students' achievement.

Regrettably, a disjunction now exists between the curricular aims we have for our students and the evidence that our high-stakes tests provide regarding whether those aims have been achieved. High-stakes tests that fail to supply per-standard feedback to parents, students, and teachers will never resolve this difficulty. What we need in this nation are high-stakes achievement tests capable of supplying, on a standard-by-standard basis for individual students, evidence of teachers' instructional successes and failures. Only then will we find that standards-based assessments can fulfill their promise to contribute to children's increased mastery of worthwhile curricular outcomes.

The need to prioritize. The second attribute of an instructionally supportive test deals with the number of content standard the test sets out to assess. Because educators hope that students will master a substantial array of worthwhile skills and knowledge, state-approved content standards often cover a vast collection of knowledge and skills. If our students actually mastered the many cognitive skills and the substantial bodies of knowledge represented in these imposing lists of state-sanctioned content standards, those students would most assuredly be well educated.

So, assuming we really do want our students to master the curricular aims that are set forth as state-approved content standards, achievement tests ought to be available that would allow educators to tell if these content standards have, in fact, been mastered by individual students. It might be supposed, of course, that the so-called "standards-based" achievement tests now used in so many of our states are already

able to provide educators with evidence about whether students have attained the curricular aspirations set forth as a state's official content standards.

Unfortunately, these "standards-based" tests typically fail to supply such evidence at a level of detail suitable for teachers' instructional decision-making. Sometimes, for example, more than 100 state-approved content standards in one subject area are supposedly measured by a 50-60 item test administered to students in about one hour. Yet, if a state's teachers receive only global test results reflecting their students' overall mastery of the state's content standards, what instructional sense can teachers make of such feedback? Which of the 100 state-approved content standards need more instruction? Which don't?

Some state educational officials, recognizing that their state's approved content standards are excessively numerous, have attempted to rectify this situation by reconceptualizing their content standards to be broader and, therefore, able to coalesce a number of more narrow content standards. Those more narrow (now-coalesced) content standards are then described as "benchmarks," "learning outcomes," "expectancies," and so on. But we usually find that only these coalesced curricular aims are stated at a level of specificity truly suitable for a teacher's instructional planning. And, of course, there are still too many of them. This curricular sleight-of-hand represents a clear instance of counterfeit coalescence.

A form of implicit instructional hypocrisy arises when a state's curricular specialists lay out a near-endless litany of content standards to be taught in the state's schools. Most experienced educators recognize that these massive arrays of content standards really represent "wish lists," that is, well-intentioned curricular wishes about what skills and knowledge educators would like students to attain. But, given the actual time available for instruction, many of those curricular yearnings simply can't come true. And it really is deceitful to suggest that these seemingly endless lists of curricular aims are actually being taught to a state's students. There simply isn't enough time available to pull off that sort of instructional miracle.

Different kinds of skills and different bodies of knowledge, of course, will require different numbers of test items in order to provide a suitable level of confidence regarding whether a *particular* skill or a *particular* body of knowledge has been mastered by a *particular* student. Generally speaking, for a test to supply per-standard results for a particular student, more test items per content standard will be needed than has typically been the case.

For certain kinds of skills, such as being able to write a persuasive essay, educators may be satisfied with one or, possibly, two student essays. After all, a student who can generate a satisfactory persuasive essay on one topic can probably (but not certainly) generate a satisfactory persuasive essay on a second topic. Yet, for other skills, such as a student's being able to calculate the areas of diverse geometric shapes, we would be uncomfortable if we were only using one or two test items.

Clearly, the number of items required will vary according to the nature of the instructional outcome being measured. But the overwhelming consequence of having a test that provides per-standard results for individual students is that *the test must assess fewer content standards*.

Because it is impossible, due to real-world time constraints, to measure all state-sanctioned content standards on a per-standard basis, the obvious way to deal with this problem is to build tests that can provide per-standard measurement of the *most important* things we want our students to master. In short, by *prioritizing* a set of content standards according to their educational importance, we can supply suitable per-standard results for the most significant content standards. It is clearly preferable to measure a modest number of content standards in a way that benefits students than it is to measure a large number of standards in a way that doesn't.

There are many considerations involved in arriving at defensible decisions about how many items are required in order to provide sensible per-standard estimates about students' mastery of a set of content standards. For instance, if a classroom teacher can reach a fairly reasonable inference about a student's mastery of a given content standard based on only a half-dozen or so items, then for purposes of that teacher's instructional decisions, the half-dozen items will suffice. However, if a student were going to be denied a grade-to-grade promotion because of the student's failure to master a specific content standard, then a half-dozen items—for that specific purpose—would surely not be enough for such a significant decision.

Clearly, the more items the better—for purposes of determining a student’s mastery of just about any content standard. But the real world of education rarely presents us with that kind of “more-items” option. Thus, we must usually cope with the inherent volatility of students’ test performances by employing the most appropriate number of items that, based on practical realities, we can employ. For different kinds of educational decisions, we will sometimes need to aggregate students’ performances on sets of items based on different content standards.

To illustrate, if results of a 50-item achievement test were to be used as one of several criteria in deciding whether a student should be advanced to the next grade level, it might be necessary to aggregate all items from the test that was attempting to assess five content standards with 10 items per standard. Yet, for a teacher’s instructional purposes, the 10-items for each content standard might supply sufficiently adequate insights about the effectiveness of the teacher’s instruction aimed at each of the five standards being assessed. In other words, the aggregation of items for purposes of making inferences about students’ status depends on the nature of the inference involved and, most critically, on the decision that will subsequently be based on that inference.

The trade-off in this instance hinges on the desire to assess a greater number of high-import content standards, yet do so with sufficient numbers of items to provide reasonably accurate per-standard estimates of a students’ status—and to do so while

avoiding the use of excessively lengthy tests. As indicated above, the best way to cope with this problem is to employ item-aggregation procedures consonant with the test-based decisions to be made. In short, the gravity of the decision-at-issue should play the chief role in determining numbers of items as well as item-aggregation procedures.

The prioritization of a state's content standards, of course, will leave many praiseworthy curricular aims unassessed by a state's high-stakes test. But those unassessed content standards, of course, will often be genuinely important. We want our students to master those content standards even if those curricular aims are not measured by a statewide high-stakes test.

One way to support teachers in their efforts to promote students' mastery of those content standards unassessed by a high-stakes statewide tests, is to make suitable assessment tools available to teachers for their on-going, in-classroom assessment. Teachers must also be provided with sufficient staff-development programs so that a state's teachers can construct their own classroom assessments related to key state-unassessed content standards.

Because of the instructionally supportive nature of the sorts of high-stakes tests described here, a state's teachers will often be able to successfully *and efficiently* promote students' mastery of the prioritized content standards measured by such an instructionally supportive test. This instructional efficiency will often make available sufficient instructional time for teachers to deal instructionally with many of the curricular

aims that, although important, are not currently assessed by a statewide high-stakes tests. The availability of optional classroom assessments measuring state-unassessed content standards will also help.

Teacher-friendly assessment descriptions. Finally, the third attribute of an instructionally supportive test revolves around the clarity with which a test describes what it is designed to measure. Statements of curricular aims, all by themselves, rarely communicate with sufficient clarity what is cognitively sought from students. Yet, if teachers do not truly understand the nature of what a particular content standard calls for from students, those teachers will have a difficult time devising and delivering effective instruction aimed at such an ill-defined content standard.

This situation can be rectified if an instructionally supportive standards-based test is accompanied by a lucid, instructionally oriented description of the cognitive demand(s) imposed on students by *each* content standard assessed. These *assessment descriptions* should capture the instructionally relevant essence of the curricular outcome being assessed, but should describe that essence in a concise, teacher-palatable manner. These assessment descriptions must be used volitionally by busy teachers, so they need to be supplied in a few brief, plain-talk paragraphs.

The heart of a well-formed assessment description is its succinct isolation of the cognitive demand(s) required if students are going to display mastery of a particular content standard. If there are key enabling subskills or bodies of knowledge that

students must possess in order to respond satisfactorily to the cognitive demand(s) spelled out in an assessment description, then such enabling knowledge and subskills should also be identified.

Once a teacher becomes familiar with an assessment description and the content standard on which that description is based, the teacher should be able to design a more effective instructional sequence in order to promote students' mastery of the content standard being measured. Suppose, for example, that several key enabling subskills and bodies of knowledge have been identified in an assessment description as precursive to a student's ultimate skill-mastery. Then the student must either already possess such enabling skills and knowledge or, if not present, the student must be taught to master those precursors of overall skill-mastery.

In addition to the assessment description's verbal depiction of the key cognitive demand(s) sought, as well as any necessary enabling subskills and/or knowledge, each assessment description should be accompanied by several *illustrative but nonexhaustive sample items*. Such items, when presented along with the assessment description, will assist teachers understand more fully what they must promote in order for students to attain skill-mastery of the content standard being tested. If possible, the illustrative items should be varied in nature (thus indicating that a student's skill-mastery should be sufficiently generalizable so that it can satisfactorily cope with *different* kinds of test items). An exception to the requirement for varied illustrative item-types would arise in the atypical instance for which only a single item-type is appropriate, for

instance, students' writing samples. Even in those one item-type situations, however, several illustrative items should still be appended to the appropriate assessment description.

The function of the assessment descriptions and their accompanying sample items is to communicate to teachers in what ways a given content standard might be assessed by a high-stakes test. But that communication, to be truly successful, must make instructional sense. The nature of what is to be measured, in a satisfactory assessment description, will have been spelled out clearly by individuals who possess not only instructional acumen, but also ample classroom experience. Assessment descriptions delineate a test's cognitive demand, but do so from an *instructional* perspective.

A Different Kind of High-Stakes Test

Three key attributes of an instructionally supportive standards-based test have been described here, namely, (1) per-standard reporting, (2) assessment of only highest-priority content standards, and (3) the provision of instructionally illuminating assessment descriptions. And, even though an instructionally supportive test can serve as a potent catalyst for improved instruction, it can still provide to accountability programs the kind of evaluative data they must have to determine which schools or which districts have been successful. And which ones haven't.

But, assuming that educational accountability is here to stay, why should American educators not employ accountability tests that nurture children's learning instead of negating it? Carefully crafted instructionally supportive tests can serve as such accountability tests.

W. James Popham is an Emeritus Professor in the UCLA Graduate School of Education and Information Studies.