

STANDARDS-BASED ASSESSMENT: SOLUTION OR CHARADE?*

W. James Popham
University of California, Los Angeles

Much of today's large-scale educational testing is described as "standards-based assessment." Such labeling stems directly from our national emphasis on promoting students' mastery of *content standards*, that is, the knowledge and skills those students are supposed to learn in school. So, if the curricular goals of the nation's educators are to be identified as "standards," it follows logically that assessments designed to tell if such goals have been achieved should be described as "standards-based."

A Key Reform-Strategy Component

Standards-based *assessment*, in fact, has been widely touted as a catalytic measurement strategy—a strategy that can substantially bolster the standards-based *reform* efforts currently encountered in so many states. The results of standards-based assessments, it is argued, will allow educators and other concerned clienteles to determine whether a state's schools have been successful in promoting students' mastery of that state's content standards.

A state's reform efforts, it is claimed, will be more effective because standards-based assessments can show the world whether students are actually mastering the outcomes they are supposed to master. Successful school-level programs can be emulated; unsuccessful programs can be fixed. Standards-based assessment, therefore, becomes a key component in an overall game-plan to improve educational quality. Proponents of standards-based assessment contend that it will spur educators toward increasingly successful instruction.

As standards-based assessment is currently being carried out in most settings, however, I think its instructional payoffs are illusory. With few exceptions these days, standards-based assessment makes little, if any, contribution to the quality of education that a state's children receive. Standards-based assessment, in my view, is typically a fine-sounding but feckless charade.

What is Standards-Based Assessment?

The label, "standards-based assessment," really ought to make clear exactly what is meant by this genre of large-scale testing. But it doesn't. You see "based" can signify different things to different people. My dictionary, for example, indicates that a testing program would be standards-based only if a set of specific content standards constituted the "principal element" or the "fundamental part" of the testing program. That's a pretty stringent definition, of course, but it's what many people think about when they consider standards-based assessment, namely, an

* Presented at the annual meeting of the American Educational Research Association, Seattle, Washington, April 10-14, 2001.

assessment program in which its tests flow directly from and suitably assess a set of content standards.

Yet, to other people, a test is standards-based if it is even loosely related to a state's sanctioned content standards. Such is often the case when an off-the-shelf standardized achievement test has been adopted because it has been judged to be satisfactorily "aligned" with the state's content standards. "Alignment," however, because it can range from the rigorous to the relaxed, might best be regarded as a four-letter word. One person's completely "aligned" test is often another person's thoroughly "unaligned" test.

It is in the financial best interests of the companies that sell off-the-shelf standardized achievement tests to perceive all sorts of alignment between their tests' items and the content standards being promoted in a state where their tests might be adopted. And even in states where a determination of alignment is made by educators in that state (rather than by members of the test publisher's staff), we often see excessively generous estimates of tests-versus-standards alignments.

It is certainly true that if a state's officials judge the nationally marketed standardized achievement tests on a comparative basis, one of them will be best aligned with that state's content standards. But is that "best alignment" sufficient for the test to stimulate improved standards-related instruction?

In many states, of course, tests will have been custom-built to reflect the state's content standards. One would think that such customized tests would, indeed, do a better job in lining up with the state's content standards. And they probably do. That lining-up job, however, may still not be good enough.

Problems with Customized Standards-Based Tests

A critical shortcoming of most state's customized standards-based tests can be traced directly to the state's content standards. Those standards are typically way too numerous and way too vague. The large number of content standards being promoted in many states poses an insurmountable obstacle for test-developers who try to measure those content standards with assessments taken by children during typical test-administration periods.

In many settings, for example, the state's content standards are described at a level of generality too broad for a state's educators to tackle instructionally. As a result, each standard is sometimes divided into *benchmarks*, *performance indicators*, or some other subcategory closer to the "chunk size" teachers think about when they plan their instruction. Almost invariably, however, the numbers of these subcategorized knowledge and skills are overwhelmingly large. Those numbers are too large not only for teachers to manage intellectually, but for testers to assess in the time typically allocated for testing. As a consequence, neither teachers nor parents are given adequate feedback at *an appropriate level of specificity*, that is, at a level of descriptive clarity sufficient for anyone to take any meaningful instructional action.

In some states there is only *a single score* provided in, say, language arts. That single score is supposed to represent a child's overall mastery of a substantial number of generally stated content standards—*each* of which often subsumes a large number of subcategorized outcomes. How can a *single* score provide any illumination regarding the *specific* sorts of things that children are or aren't learning? Unless a standards-based test provides a state's teachers with some sort of indication of how well students are learning the *specific* skills or knowledge that teachers can address instructionally, the test will do little to further the quality of instruction.

Moreover, because students' scores on a customized standards-based test are often seen as a reflection of how effective the state's educators have been, it will often be the case that no progress will be indicated by such test-scores. After all, the state's teachers will be overwhelmed by the many, many things they are supposed to teach. The original architects of the state's content standards will have identified "all the good things" they would like children to learn. But in that process they will have created curricular targets far too numerous to be of any practical value. The state's teachers simply won't know where to aim their instruction. There's just too much to teach. And standards-based assessments won't remedy this situation because the relationship between multitudinous standards and a test's actual items will be thoroughly murky.

Summing up, standards-based assessments that provide only single scores, or even subcategory scores, at a level of generality too broad for instructional-planning decisions will fall far short of their potential. Are such tests better than no tests at all? Of course they are. Are tests somewhat aligned with a state's content standards better than tests less well aligned to those standards? Of course they are.

But my disappointment in today's standards-based assessments is that they could be doing so much more. Remember, these assessments are touted as an integral feature of an overall standards-rooted reform strategy to improve instructional quality. Yet, other than providing a general and often uninterpretable picture of students' standards-status, most of today's standards-based measurements misfire from an instructional perspective. Their contribution to improved instruction is almost undetectable.

Instruction-Enhancing Assessments

And that's what I'd like to describe: how to build standards-based assessments that will serve as catalysts to improve instruction. This description will be brief. I've tried to provide more detailed, step-by-step guidance for building such tests elsewhere (Popham, 2001). I will now present four rules which, if followed, would lead to the generation of standards-based assessments that make meaningful contributions to improved instructional quality. Although I will not elaborate extensively on each rule, it should be clear that the kind of standards-based tests I have in mind differ decisively from the sorts of standards-based tests now common in the U.S.

Prioritizing Outcomes

To create a standards-based test that stimulates improved instruction, the first rule calls for us to avoid the assessment hypocrisy that a single standards-based test can do a satisfactory job in assessing a laundry list of content standards. Any measurement specialist who claims a 60-minute high-stakes test can accurately assess 30-50 educational outcomes is embroidering psychometric truth. And that recognition leads directly to my first rule:

Rule 1. Require curricular personnel to prioritize the most important outcomes they want children to achieve, then develop tests to assess only the highest priority outcomes that can be both accurately assessed and instructionally accomplished.

This initial rule is intended to counteract the well-intentioned but instructionally counterproductive bent of curricular specialists to embrace too much in their content standards. Elsewhere I have referred to such litanies of knowledge and skills as *wish-list content standards* because they represent all the good things that curriculum specialists yearn to see achieved in their fields (Popham, 2000). But wishing won't make it so. And Rule 1 requires that the curriculum folks in authority set about to ruthlessly *prioritize* the skills or the bodies of knowledge they'd like to see children in their state master.

Here's how it might work. A group of state-authorized curricular specialists in a given subject would review all of that subject's content standards under consideration, then use group-based *ratings* to split the standards into three groups, for instance, absolutely essential, highly desirable, and desirable. *All* of these content standards can still be promoted instructionally in the state's classrooms. Then, however, the curriculum specialists would be required to focus *only* on the "absolutely essential" standards. At that point those outcomes must be *ranked* from most to least important. It will be a difficult assignment, but curriculum personnel (screaming, to be sure) can do it.

The measurement crowd then swoops in to design standards-based assessments, but only for a certain number of highest ranked outcomes. The outcomes that will be assessed are those that, in the test-administration time available, can be measured so a valid inference can be drawn about students' mastery. Accurate *per-standard* information about students' status must be available from an instructionally helpful standards-based test. Because of the test-administration time periods typically available, the result is usually a standards-based test in a given subject area that will assess only a half-dozen or so content standards rather than 20-30 such standards.

Moreover, the content standards to be assessed must identify knowledge or skills that, realistically, can be accomplished instructionally by any reasonably effective teacher. To make sure that both these conditions have been satisfied (accurate assessment and instructional accomplishability), the test's developers will find that they must approach their work from both a testing and teaching perspective. Psychometric purists won't be able to follow Rule 1 unless they get help from people who possess instructional acumen.

Rule 1, of course, does not preclude a state's teachers from using their own classroom assessments to measure students' attainment of those content standards that, because of the prioritizing operation, failed to emerge from the final assessment-cut. Indeed, if Rule 1 were followed, a state's teachers would need ample encouragement and staff-development support so they could assess their own students' mastery of other worthwhile content standards.

What Rule 1 assumes, of course, is that it is impossible to assess with accuracy and instructional facilitation *all* of the content standards that we want children to learn. Rule 1 simply calls for assessment honesty. Without the prioritizing of content standards, large-scale tests will make little, if any, contribution to instructional improvement.

Conceptualizing Assessment Tasks Instructionally

If Rule 1 were to be followed, a standards-based test would be focused only on a modest number of high-import instructional outcomes. Then the *tasks* to which students must respond in any such test must be of the sort indicated in Rule 2 below:

Rule 2. Construct all assessment tasks so an appropriate response will typically require the student to employ (1) key enabling knowledge and/or subskills, (2) the evaluative criteria to be used in judging a response's quality, or (3) both.

This second rule calls for test developers to deliberately conceptualize the tasks (items, prompts, etc.) to which students must respond so that a successful (correct) response to those tasks will require a student to utilize the elements that are at the heart of a student's satisfactory achievement of the content standard.

To do this, a test-developer must approach the task-construction operation from a decisively *instructional* perspective. What this second rule attempts to do is lay out assessment tasks in such a way that teachers will be inclined to address instructionally a task's key ingredients. The assessment tasks, in essence, are built so that while they may not *drive* teachers in a particular direction (as in "measurement-driven instruction"), they are surely intended to *direct a teacher's attention* toward the instructionally salient features of those tasks.

Rule 2 does not force teachers to teach in a particular fashion, of course. There are numerous highways to Instructional Mecca. However, Rule 2 waves what's instructionally important squarely in the teacher's face, and says, "Pay attention!"

Supplying Assessment Descriptions

A standards-based test that will have a positive impact on instruction will be one that *clarifies* what teachers are expected to accomplish. Standards-based assessments should, in a technical sense, *operationalize* what a particular content standard actually means. And that's where Rule 3 comes in. Its intention is to help teachers get a solid handle on what a standards-based test is actually measuring:

Rule 3. Create a sufficiently clear description of the knowledge and/or skills represented by the test so that teachers will have an understanding of the cognitive demands required for students' successful performance.

Test developers must create an *assessment description* for each content standard being measured on a standards-based test. That assessment description must spell out the nature of the *cognitive demands* imposed on students by each measured content standard. Thus, if a standards-based test were built to measure five highest-priority outcomes, there would be five assessment descriptions for that test. Each description, based on the content standard it addresses, would attempt to lay out in a concise, teacher-palatable fashion, just what it is that a student must do in order to successfully attain the content standard involved.

Because these assessment descriptions are intended to help teachers better understand the pivotal *instructional* elements of a content standard, it is imperative that the descriptions be short and easily read. Busy teachers aren't going to wade through an assessment encyclopedia. More than a little descriptive artistry will be required to convey a content standard's chief cognitive demands with fidelity, yet in a sufficiently succinct form so that teachers will be willing to use the resulting description.

The use of illustrative items often proves most helpful in giving life and lucidity to an assessment description. But, in order to promote the student's *generalizable* mastery of a content standard, the sample items accompanying an assessment description should be both *diverse* in nature and *non-exhaustive*. Teachers must recognize that the only way their students will become able to display attainment of a content standard will be to achieve flat-out, generalizable mastery of the standard—no matter how it is assessed!

Making Certain

The final rule, as you will see, is less of a test-construction rule and deals more with making sure that a standards-based test is, in fact, a winner:

Rule 4. The items and description(s) of any high-stakes test should be reviewed at a level of rigor commensurate with the intended uses of the test.

Rule 4 calls for a high-stakes test, such as a statewide assessment employed for evaluating schools, to be scrutinized according to the intended consequences of the test's use. I refer not only to the items on the test, but also to the assessment description(s) accompanying that test.

The kind of item judgments I have in mind would deal with (1) an item's congruence with the content standard it is supposed to help assess, (2) the degree to which effective instruction directed toward an item's content standard is apt to allow students to respond to the item correctly, (3) the freedom of an item's content from dominant influence by a student's socioeconomic status or inherited academic aptitudes, and (4) the absence of biased content that might offend or unfairly penalize students on the basis of personal characteristics such as race or

gender. For the assessment descriptions, I think each should be reviewed both for its instructional illumination and for its palatability (language and brevity) to teachers.

The individuals who would supply these reviews should, in the main, be classroom teachers. However, curricular or instructional specialists could also take part in the review operations. All reviews should be devised and implemented by individuals other than the agency that developed the standards-based tests being reviewed. Elsewhere, I have spelled out possible language to use in the questions for reviewing items and assessment descriptions (Popham, 2001).

Educational Assessments

Educational assessment is getting a bad reputation—among educators. That negative image stems from the almost universal preoccupation with test-based accountability on the part of the public as well as the educational measurement community itself. Yesterday’s high-stakes tests have become today’s high-stakes tests, surely to be followed by tomorrow’s even higher-stakes tests. Everyone wants hard evidence showing whether the quality of schooling is good or bad. And educational assessment has been chosen as the vehicle to supply such evidence.

But as far as most of the nation’s classroom teachers are concerned, the evidence emanating from the use of educational assessments is typically worthless. More often than not, the evidence is also depressing. Students’ test scores, from year to year, rarely move dramatically higher. As a result, educational assessors characteristically are the bearers of bad tidings. And history has shown us that messengers who carry negative news are rarely greeted with jubilation.

One reason today’s educational tests fail to detect substantial year-to-year improvements in students’ scores is that the tests have been constructed chiefly for an *accountability* mission rather than an *instructional* function. With so much high-level clamor for evidence of educational success (or lack of it) these days, we should not be surprised that assessment specialists have devoted themselves chiefly to the generation of psychometrically sound accountability tests. Yet, as a consequence of that preoccupation with the accountability function of assessment, we should be equally unsurprised that educators regard today’s tests largely as “no-win” measuring sticks.

Yet, educational assessments should not be tombstones to commemorate instructional impotence. Rather, even large-scale assessments can be crafted so they help *educate* our students. Today’s popular standards-based reform programs can be made more successful if the assessments on which they rely can be made more instructionally influential. Standards-based assessments *can* be part of the solution. But they must be built differently.

References

- Popham, W. James (2001). *Educational Assessment: High Quality Testing for a High Stakes World*. Alexandria, VA: Association for Supervision and Curriculum Development.
- Popham, W. James (December 2000). "Assessing Mastery of Wish-List Content Standards." *The NASSP Bulletin*, 84, 620: 30-36.