Putting the Standardized Test Debate in Perspective

When used correctly, standardized tests *do* have value, but they provide only part of the picture and have limits—which we must understand and work to improve.

Are the criticisms of educational testing valid, or do most of the objections stem from the fact that such tests are often misused? By far the most common type of standardized test is the norm-referenced test—that in which a student's performance is systematically compared with the performance of other (presumably) similar students. Minimum competency and criterion-referenced tests—those that measure student performance against established criteria—can also be standardized. However, not coincidentally, most criticism has been leveled at standardized, norm-referenced tests.

Criticisms of Standardized Tests

Among the current criticisms, a few stand out as most pervasive and most bothersome to those who worry over whether to support or oppose standardized testing. In this article, we'll look at seven of the most common criticisms.

Criticism #1: Standardized achievement tests do not promote student learning. Critics charge that standardized achievement tests provide little direct support for the "real stuff" of education, namely, what goes on in the classroom. They do nothing, critics contend, to enhance the learning process, diagnose learning problems, or provide students rapid feedback.

True, standardized tests do paint student performance in broad brush strokes. They provide general performance information in content areas like math or reading—as the test developers have defined these areas. They do not, nor are they *meant* to, pick up the nuances of performance that characterize the full range of a student's skill, ability, and learning style. Of course, we hope that standardized test results are only a small portion of the assessment information a teacher relies on in making academic decisions about students or curriculum. Good classroom assessment begins with a teacher's own observations and measurement of what students are gaining from instruction every day. Standardized testing can never replace that teacher-centered assessment. But it *can* supplement it with additional information that may help clarify a larger picture of student performance.

Criticism #2: Standardized achievement and aptitude tests are poor predictors of individual students' performance. While some tests may accurately

Photograph by Harriet Sutherland

Standardized tests were never intended to assess the diverse array of learning that occurs daily in classrooms, but they can paint a valuable portrait of students' performance in broad terms.

predict future performances of *groups*, critics of testing argue that they are often inaccurate predictors of *individual* performance. Remember Einstein flunked 6th grade math, the critics point out eagerly. Clearly, no test can tell everything. If standardized tests were thousands of items long and took days to administer, they'd probably be better predictors than they are now. But remember—there are predictions and predictions. When a person passes a driver's test, we can't say she'll never speed or run a red light. Similarly, when a child scores well on a standardized reading test, that doesn't mean we can kick back and say, "Well, he's a terrific reader, all right. That's how it will always be." Ridiculous. Maybe he felt extra confident. Maybe the test just happened to touch on those things he knew well. But if we look at *all* the students with high scores and *all* those with low scores, we can safely predict more reading difficulties among students with low scores.

What all this means is that in a standardized test we have the best of one world—a measure that is relatively accurate, pretty good at what it does, but necessarily limited in scope.

Because there are so many drivers to be tested and only a finite amount of time, we cannot test each driver in every conceivable driving situation; and, similarly, we cannot measure all we might like to measure about a child's reading skills without creating a standardized test so cumbersome and complex no one would want to use it. The world of testing is, to a large extent, a world of compromise.

Criticism #3: The content of standardized achievement tests is often mismatched with the content emphasized in a school's curriculum and classrooms. Because standardized tests are intended for broad use, they make no pretense of fitting precisely and equally well the specific content being taught to 3rd graders in Salt Lake City's public schools and their counterparts at the Tickapoo School downstate. Instead, they attempt to sample what is typically taught to *most* 3rd graders in *most* school districts. The result is a test that reflects most curriculums a little, but reflects none precisely. For most users, there are big gaps—whole lessons and units and months of instruction skimmed over or left out altogether. Or the emphasis may seem wrong—too much attention to phonics, not enough on reading for meaning, perhaps. Again, the problem is the size of the test. We simply cannot cover in 10 or 20 test items the richness and diversity that characterize many current curriculums.

Educators have a serious ethical obligation to use tests *well*, if we use them at all.

Criticism #4: Standardized tests dictate or restrict what is taught. Claims that standardized tests dominate school curriculums and result in "teaching to the test" are familiar and can be leveled at any type of standardized testing that has serious consequences for the schools in which it is used. On the surface it may seem inconsistent to claim that standardized tests are mismatched with what is taught in the schools and at the same time to complain that the tests "drive the curriculum." But those two allegations are not necessarily at odds. The first is grounded in a fear that in trying to represent everyone somewhat, standardized tests will wind up representing no one really well; the second arises from the consequent fear that everyone will try to emulate the generic curriculum suggested by the test content. This doesn't have to happen, of course.

Further, to the extent it does happen, it seems absurd to blame the test. The question we really need to be asking is "How are decisions about curriculum content being made?" There's often considerable fuzziness on that issue. Here's one sobering note:

> Achievement test batteries are designed around what is thought to be the content of the school curriculum as determined by surveys of textbooks, teachers, and other tests. Textbooks and curriculums are designed, on the other hand, in part around the content of tests. One cannot discern which side leads and which follows; each side influences the other, yet nothing assures us that both are tied to an intelligent conceptualization of what an educated person ought to be.[1]

Criticism #5: Standardized achievement and aptitude tests categorize and label students in ways that cause damage to individuals. One of the most serious allegations against published tests is that their use harms students who are relentlessly trailed by low test scores. Call it categorizing, classifying, labeling (or mislabeling), or whatever, the result is the same, critics argue: individual children are subjected to demeaning and insulting placement into categories. The issue is really twofold: (1) tests are not infallible (students can and do change and can also be misclassified); and (2) even when tests *are* accurate, categorization of students into groups that carry a negative connotation may cause more harm than any gain that could possibly come from such classification.

Published tests, critics claim, have far too significant an effect on the life choices of young people. Some believe that achievement and intelligence tests are merely convenient and expedient means of classifying children and, in some cases, excluding them from regular education. But here again, it's important to raise the question of appropriate use. Even if we agree that it's okay to classify some children in some cases for some purposes, we must still ask whether standardized tests provide sufficient information to allow for intelligent decisions. We must also ask whether such tests provide any really useful information not already available from other sources.

Here's something to keep in mind, too. Some test results rank students along a percentile range. For instance, a student with a percentile ranking of 75 on a reading test may be said to have performed better than 75 percent of the other students who took the same test. But a difference in performance on even *one test item* could significantly raise or lower that

Photograph by Betty Buchignani

A teacher's insights into her students' learning and the results of standardized tests, properly used, together provide much more than either could provide alone.

percentile ranking. Knowing this, should we classify students on the basis of standardized tests? That probably depends on the consequences, on whether the information is appropriate and sufficient for the decision at hand, and on whether there is any corroborating evidence. Suppose we identify talented and gifted students on the basis of standardized math and reading tests. We ought, then, to at least be able to show that high performance on those tests is correlated directly with high probability of success in the talented and gifted program.

Criticism #6: Standardized achievement and aptitude measures are racially, culturally, and socially biased. Perhaps the most serious indictment aimed at both norm-referenced and minimum competency tests is that they are biased against ethnic and cultural minority children. Most published tests, critics claim, favor economically and socially advantaged children over their counterparts from lower socioeconomic families. Minority group members note that many tests have disproportionately negative impact on their chances for equal opportunities in education and employment. We must acknowledge that even well-intentioned uses of tests can disadvantage those unfamiliar with the concepts and language of the majority culture producing the tests. The predictable result is cultural and social bias—failure of the test to reflect or take into account the full range of the student's cultural and social background.

A conviction that testing is biased against minorities has led some critics to call for a moratorium on testing and has also prompted most of the legal challenges issued against minimum competency tests or the use of norm-referenced standardized tests to classify students. It is tempting, in the face of abuses, to outlaw testing. But simplistic solutions rarely work well. A more conservative, and far more challenging, solution is to improve our tests, to build in the sensitivity to cultural differences that would make them fair for all—and to interpret results with an honest awareness of any bias not yet weeded out.

Making such an effort is crucial, if one stops to consider one sobering thought. Assume for the moment that there *is* a bit of cultural bias in college entrance tests. Do away with them, right? Not unless you want to see college admission decisions revert to the still more biased "Good Old Boy" who-knows-whom type of system that excluded minorities effectively for decades before admissions tests, though admittedly imperfect, provided a less biased alternative.

Criticism #7: Standardized achievement and aptitude tests measure only limited and superficial student knowledge and behaviors. While test critics and supporters agree that tests only sample whatever is being tested, critics go on to argue that even what is measured may be trivial or irrelevant. No test items really ask "Who was buried in Grant's Tomb?" but some are nearly that bad.

They don't have to be. The notion that multiple choice tests can tap only recall is a myth. In fact, the best multiple choice items can—and do—measure students' ability to analyze, synthesize information, make comparisons, draw inferences, and evaluate ideas, products, or performances. In many cases, tests are improving, thanks in large part to critics who never give up.

Better Than the Alternatives

No test is perfect, and taken as a whole, educational and psychological measurement is still (and may always be) an imperfect science. Proponents of standardized tests may point to psychometric theory, statistical evidence, the merits of standardization, the predictive validity of many specific tests, and objective scoring procedures as arguments that tests are the most fair and bias-free of any procedures for assessing learning and other mental abilities. But no well-grounded psychometrician will claim that tests are flawless, only that they are enormously useful.

What do they offer us that we couldn't get without them? Comparability, for one thing. Comparability in the context of the "big picture," that is. It isn't very useful, usually, for one teacher to compare his or her students' performance with that of the students one room down and then to make decisions about instruction based on that comparison. It's too limited. We have to back away to get perspective. This is what standardized test results enable us to do—to back off a bit and get the big, overall view on how we can answer global questions: In *general*, are 3rd graders learning basic math? Can 6th graders read at the predefined level of competency?

Thus, such tests will be useful to us if we use them as they were intended and do not ask them to do things they were never meant to do, such as giving us a microscopic view of an individual student's range of skills.

Appropriate Use Is the Key

On their own, tests are incapable of harming students. It is the way in which their results can be misused that is potentially harmful. Critics of testing of-

ten overlook this important distinction, preferring to target the instruments themselves, as if they were the real culprits. That is rather like blaming the hemlock for Socrates' fate. It is palpable nonsense to blame all testing problems on tests, no matter how poorly constructed, while absolving users of all responsibility—not that bad tests should be condoned, of course. But even the best tests can create problems if they're misused. Here are some important pitfalls to avoid.

1. *Using the wrong tests.* Schools often devise new goals and curriculum plans only to find their success being judged by tests that are not relevant to those goals or plans yet are imposed by those at higher administrative levels. Even if district or state level administrators, for example, have sound reasons for using such tests at *their* level, that does not excuse any school for allowing such tests to be the *only* measures of their programs. Teachers and local administrators should exert all the influence they can to see that any measures used are appropriate to the task at hand. They can either (1) persuade higher administrators to select new standardized achievement or minimum competency measures that better match the local curriculum or (2) supplement those tests with measures selected or constructed specifically to measure what the school is attempting to accomplish.

Subtle but absurd mismatches of purpose and test abound in education. Consider, for instance, use of statewide minimum competency tests to make interschool comparisons, without regard for differences in student ability. Misuse of tests would be largely eliminated if every test were carefully linked with the decision at hand. And if no decision is in the offing, one should question why *any* testing is proposed.

2. *Assuming test scores are infallible.* Every test score contains possible error; a student's *observed* score is rarely identical to that student's *true* score (the score he or she would have obtained had there been no distractions during testing, no fatigue or illness, no "lucky guesses," and no other factors that either helped or hindered that score). Measurement experts can calculate the probability that an individual's *true* score will fall within a certain number of score points of the *obtained* score. Yet many educators ignore measurement error and use test scores as if they were highly precise measures.

On their own, tests are incapable of harming students. It is the way in which their results can be misused that is potentially harmful.

3. *Using a single test score to make an important decision.* Given the possibility of error that exists for every test score, how wise is it to allow crucial decisions for individuals (or programs) to hinge on the single administration of a test? A single test score is too suspect—in the absence of supporting evidence of some type—to serve as the sole criterion for *any* crucial decision.

4. *Failing to supplement test scores with other information.* Doesn't the teacher's knowledge of the student's ability count for anything? It should. Though our individual perceptions as teachers and administrators may be subjective, they are not irrelevant. Private observations and practical awareness of students' abilities can and should supplement more objective test scores.

5. *Setting arbitrary minimums for performance on tests.* When minimum test scores are established as critical hurdles for selection and admissions, as dividing lines for placing students, or as the determining factor in awarding certificates, several issues become acute. Test validity, always important, becomes crucial; and the minimum standard itself must be carefully scrutinized. Is there any empirical evidence that the minimum standard is set correctly, that those who score higher than the cutoff can be predicted to do better in subsequent academic or career pursuits? Or has the standard been set through some arbitrary or capricious process? Using arbitrary minimum scores to make critical decisions is potentially one of the most damaging misuses of educational tests.

6. *Assuming tests measure all the content, skills, or behaviors of interest.* Every test is limited in what it covers. Seldom is it feasible to test more than a sample of the relevant content, skills, or traits the test is designed to assess. Sometimes students do well on a test just because they happen to have read the *particular* chapters or studied the *particular* content sampled by that test. Given another test, with a different sampling of content from the same book, the students might fare less well.

7. *Accepting uncritically all claims made by test authors and publishers.* Most test authors or publishers are enthusiastic about their products, and excessive zeal can lead to risky and misleading promises. A so-called "creativity test" may really measure only verbal fluency. A math "achievement" test administered in English to a group of Inuit Eskimo children (for whom English is a second language) may test understanding of English much more than understanding of math.

8. *Interpreting test scores inappropriately.* The test score *per se* tells us nothing about *why* an individual obtained that score. We watched the SAT scores fall year after year, but there was nothing in the scores themselves to tell us *why* that trend was downward. There turned out, in fact, to be nearly as many interpretations of the trend as there were interpreters.

A student's test *score* is not a quali-

tative evaluation of performance, but rather, a mere numeric indicator that lacks meaning in the absence of some criteria defining what constitutes "good" or "bad" performance.

9. *Using test scores to draw inappropriate comparisons.* Unprofessional or careless comparisons of achievement test results can foster unhealthy competition among classmates, siblings, or even schools because of ready-made bases for comparisons, such as grade-level achievement. Such misuses of tests not only potentially harm both the schools and the children involved, but also create an understandable backlash toward the tests, which should have been directed toward those who misused them in this way.

10. *Allowing tests to drive the curriculum.* Remember that *some* individual or group has selected those tests, for whatever reason. If a test unduly influences what goes on in a school's curriculum, then someone has allowed it to override priorities that educators, parents, and the school board have established.

11. *Using poor tests.* Why go to the effort of testing, then employ a poorly constructed or unreliable measure—especially if a better one is at hand? Tests can be flawed in a multitude of ways, from measuring the wrong content or skills (but doing it well) to measuring the correct content or skills (but doing it poorly). Every effort should be made to obtain or construct the best possible measures.

12. *Using tests unprofessionally.* When educational tests are used in misleading or harmful ways, inadequate training of educators is often at fault. When test scores are used to label children in harmful ways, the fault generally lies with those who affix the labels—not with the test. When scores are not kept confidential, that is the fault of the person who violated the confidence, not the test maker. In short, as educators, we have a serious ethical obligation to use tests *well*, if we use them at all.

In Search of a Balanced View

Not all criticisms of tests can be deflected by claiming that they merely reflect misuses of tests. There are also apparent weaknesses in many tests, partly because we have yet a good deal to learn about measurement. We know enough already, however, to state unequivocally that uncertainty and error will always be with us, and no test of learning or mental ability or other characteristics can ever be presumed absolutely precise in its measurements. The professional judgments of teachers and other educators will continue to be essential in sound educational decision making. But we also assert—as do test advocates—that tests are often a great deal better than the alternatives. Thus, we find ourselves caught in the middle of the debate between testing critics and enthusiasts.

Tests can be flawed in a multitude of ways, from measuring the wrong content or skills (but doing it well) to measuring the correct content or skills (but doing it poorly).

The stridency of that debate occasionally calls to mind the old rhyme, "When in danger or in doubt, run in circles, scream, and shout!" In more recent years, however, there has been some softening on both sides. Measurement experts spend less time defending tests and deriding their detractors and more time working to improve the science of measurement. At the same time, they have become more comfortable in acknowledging that test scores are approximations and less obsessed with claiming unflinching scientific support for every test they devise.

Meanwhile, critics seem less intent on diagnosing psychometric pimples as terminal acne. They seem more aware that many testing problems stem from misuse, and their calls for "testing reform" have quieted somewhat as they have recognized that even the best tests, if subjected to the same sorts of misuse, would prove no more helpful. Further, most critics are beginning to acknowledge that abolishing testing would leave us with many decisions still to make—and even less defensible bases on which to make them.

But even if there are no quick-fix answers to the testing dilemma, there are things we can do. We can: (1) scrupulously avoid any misuses of tests or test results; (2) educate ourselves and our colleagues about tests so that we understand their capabilities and limitations and do not ask them to tell us more than they can; (3) stretch to the limit our creative talents in test design, teaching ourselves to develop test items that not only resound with our own thoughtful understanding of critical content but that encourage students to think; and (4) recall, even when pressed for hasty or expedient decisions, that no matter how much any test may tell us, there is always so much more to be known. □

[1]G. V. Glass, (1986), "Testing Old, Testing New: Schoolboy Psychology and the Allocation of Intellectual Resources," in *The Future of Testing*, Buros-Nebraska Symposium on Measurement and Testing, Vol. 2, p. 14, edited by B. S. Plake, J. C. Witt, and J. V. Mitchell, (Hillsdale, N.J.: Lawrence Erlbaum Associates).

Blaine R. Worthen is Professor and Chair, Research and Evaluation Methodology Program, Utah State University, Psychology Department, Logan, UT 84322. **Vicki Spandel** is Senior Research Associate, Evaluation and Assessment Program, Northwest Regional Educational Laboratory, 101 S.W. Main St., Portland, OR 97204.