Standardized achievement tests should not be used to evaluate the quality of students' schooling because there are meaningful mismatches between what is tested and what is supposed to be taught, and those mismatches are often unrecognized.

Reason 2: The Tendency to Jettison Items Covering Important Content

I attribute Reason 2 to the "Army Alpha thinking" displayed by most of today's measurement specialists. Remember, the Alpha worked well because it was able to produce a substantial degree of *score-spread* among examinees. If the Alpha's scores weren't spread out widely enough, then different recruits couldn't be contrasted with sufficient precision to distinguish between men who scored at the 83rd or 84th percentiles. The fine-grained comparisons, and the score-spread necessary to produce them, are at the heart of the Army Alpha approach to assessment, which, as noted, remains the underlying assessment approach favored by today's standardized achievement test creators.

The other factor that spurs the developers of standardized achievement tests to covet substantial score-spread is related to the technical determination of a test's *reliability*. Test publishers can compute three different kinds of reliability, all of which are rooted in the concept of *consistency*. The most commonly calculated kind of reliability refers to the consistency with which a test's items measure whatever they're measuring. But all three types of reliability will be higher if the test produces substantial score-spread. The better the score-spread, the higher the test's reliability. This occurs because the ways of calculating reliability are correlationally based, and substantial score-spread is required for high correlation—that is, reliability coefficiencies.

And why is high reliability so esteemed by the measurement staffs who create standardized achievement tests? It's simple: *High reliability sells tests*. When it comes time to select among competing standardized achievement tests, the decision makers (say, a district or state test-selection committee) will look to many evaluative factors to determine which test is best. They may give attention to

the degree of apparent alignment between published descriptions of the test's content and the locally sanctioned curricular content. The decision makers usually consider the efforts made to eliminate from the test any content that might be biased against minorities. And they always pay attention to evidence regarding the test's technical qualities, one of which is reliability.

Other factors being equal, a test that has better reliability than any of the other tests will be chosen over its competitors. Thus, test developers diligently seek high indicators of reliability and the score-spread that helps create such high reliability. Substantial score-spread not only contributes to more accurate discriminations among examinees, it also helps peddle tests. And well-peddled tests make more money for the shareholders in the corporations that build and sell standardized tests.

Now let's turn our attention to an important technical point about the nature of test items that contribute most effectively to the creation of score-spread. These are the test items that are answered correctly by roughly 50 percent of the examinees. Test folks use the term "*p*-value" to indicate the percentage of students who answer an item correctly. An item with a *p*-value of .85 would have been answered correctly by 85 percent of those who attempted to answer it. A test item answered correctly by exactly half the examinees would have a *p*-value of .50.

Items that make the best contribution to a test's score-spread are those with *p*-values in the .40 to .60 range. Not surprisingly, most items on the national standardized achievement tests have *p*-values in that range. Items that have higher *p*-values—for example, .80 or .90—are rarely included in these tests, having been determined during shakedown trials to be "too easy." Moreover, any items that produce unanticipated high *p*-values are almost always removed from a standardized test once the test is revised (typically every 5–10 years). A test item with a *p*-value of .93 just doesn't make a sufficiently substantial contribution to the production of score-spread. Indeed, an item that *all* examinees answered correctly would have a *p*-value of 1.00 and would make *zero* contribution to the creation of score-spread. Useless!

Now for the difficulty created by this relentless quest for score-spread. As it turns out, teachers tend to stress the curricular content they believe to be most important. The more significant a topic, the more likely it is that the teacher will emphasize the topic instructionally. And, of course, the more that teachers emphasize any curricular content, the better that students are likely to perform on items measuring such content. As a perverse consequence, items covering the most important things that teachers teach tend to be excluded from standardized achievement tests. Such items, even though they tap teacher-stressed content, will either not have been placed on the test to begin with, or will be discarded from the test at revision (as a consequence of high *p*-values).

Thus, the more important the content, the more likely teachers are to stress it. The more that teachers stress important content, the better that students will do on an item measuring that content. But the better that students do on such an item, the more likely it is that the item will disappear from the test. How fair is it, then, to evaluate a school's staff on the basis of a test that, because of its Army Alpha-like quest for score-spread, fails to include items covering the most important things teachers are trying to teach? And, of course, those important things will typically be the things that teachers have taught well.

The second reason, then, that standardized achievement tests should not be used to evaluate educational quality is somewhat technical, but nonetheless important:

> **Standardized achievement tests should not be used to evaluate the quality of students' schooling because the quest for wide score-spread tends to eliminate items covering important content that teachers have emphasized and students have mastered.**

These two reasons—teaching/testing mismatches and the tendency to eliminate the very items covering the most important things that teachers teach—*all by themselves* should be sufficient to disincline anyone from using students' scores on standardized achievement tests as indicators of instructional quality.

But there's a third reason, one that in my mind is far nastier. It's

the reason you'll learn about in the next chapter. And it's a reason that becomes clear only when you look carefully at the kinds of test items actually used on standardized achievement tests. Taken together, the two reasons in this chapter and the one in Chapter 4 form a potent three-point rationale for *never* using standardized achievement tests to judge educational quality.

✘ ✔ ✘ ✔ ✘