

RIGHT TASK, WRONG TOOL: EVALUATING SCHOOLS WITH STANDARDIZED TESTS

W. James Popham
University of California, Los Angeles

Most Americans, and that includes school board members, believe the best way to evaluate a school is to see how well the school's students perform on a standardized achievement test. Despite the pervasiveness of this belief, however, it is quite wrong.

Clearly, if educational policymakers are arriving at flawed conclusions about school quality by using the wrong evidence, then those conclusions are likely to produce unsound policies. Truly successful schools may be regarded as ineffective; truly unsuccessful schools may be regarded as effective. When educational policymakers are guided by the wrong measurement data, it is a certainty that misguided policies will follow.

In this brief essay I want to explain why the evaluation of educational quality using standardized achievement tests is improper. If educational policymakers begin to recognize the shortcomings of an incorrect evaluative approach, they can then move toward more defensible ways of evaluating schools.

Standardized Achievement Tests: Present and Past

There are two kinds of standardized achievement tests I'll be considering, namely, national tests and state-specific tests. Currently, there are five nationally standardized achievement tests widely used in U.S. public schools. The five nationally standardized tests, for example, *The Iowa Tests of Basic Skills*, are developed and distributed by three measurement companies. State-specific tests, such as the *Florida Comprehensive Assessment Test*, are developed on a customized basis in a hope that the test will better match a given state's curricular preferences. But these state-specific tests are often built by the same companies that develop and market the five national tests. As a consequence, state-specific tests often perform measurement tasks that are essentially identical to the measurement tasks performed by national tests.

All of these tests are *standardized* in the sense that they are to be administered, scored, and interpreted in a standard, predetermined manner. They are *achievement* tests rather than *aptitude* tests because an achievement test is supposed to assess students' knowledge and skills. Aptitude tests, such as the SAT and ACT, are intended to predict how well a high school student will perform in a subsequent academic setting, for instance, in college. I will be focusing here, however, only on achievement tests.

The ancestry of today's standardized achievement tests can be traced directly back to World War I when, to assist the U.S. Army in identifying prospective officer-training candidates, the *Army Alpha* was created. The *Alpha*, a group-administered

* A version of this analysis appeared in the February 2002 issue of the *American School Board Journal*, Vol. 189, No. 2, pp. 18-22.

intelligence test administered to nearly two million Army recruits, was designed to sort out examinees based on their relative mental abilities. The test performed its measurement mission with striking success.

Because the *Alpha* worked so well, it became the prototype for almost all subsequent educational tests built in the U.S., both achievement tests as well as aptitude tests. Today's standardized educational achievement tests are patterned after the *Alpha* and attempt to carry out its measurement mission of providing *relative* score-based comparisons of examinees.

Incompatible Measurement Missions

When we evaluate schools, our chief concern should be on determining the quantity and quality of what students have learned there. Unfortunately, a test whose overriding purpose is to provide relative comparisons among test-takers can never be a suitable tool for evaluating what students have learned in school. Here's why.

In order for a standardized achievement tests to provide fine-grained comparisons such as, "Jill scored at the 86th percentile," or "Bill scored at the 88th percentile," the test must produce sufficient *score-spread*. Score-spread refers to the range of different scores a test typically provides. If a standardized test's scores are too "bunched up," that is, if there is insufficient score-spread, then the test will not permit the precise score-contrasts that, since the days of the *Army Alpha*, have been the mainstay of educational achievement testing in this nation.

Because the administration time allowed for standardized achievement testing is often only about an hour (otherwise, students become rebellious), it is imperative for the developers of these tests to choose items, about 50 or 60, that produce adequate score-spread.

Thus, suppose a test-developer is faced with a choice between (1) *Item A* measuring something really important that students should be taught or (2) *Item B* that contributes meaningfully to the production of score-spread. For traditionally constructed standardized achievement tests, *Item B* will almost certainly be selected. The test-developers' relentless quest for score-spread, in fact, leads to the creation of standardized achievement tests that do a dismal job of measuring how much students have learned in school.

What's Learned in School

As indicated above, the most important factor in evaluating the success of a given school's staff should surely be how much the school's students have learned. But this is not what's being measured by a traditional standardized achievement test. In fact, the technical test-development procedures employed to build these tests tend to make them remarkably *insensitive* to the detection of the things students have learned

in school, even in a highly effective school. Remember, assessing what students have learned in school is *not* the measurement mission of any *Alpha*-sired achievement test.

To illustrate how instructional insensitivity gets incorporated in a traditionally constructed standardized achievement test, consider the nature of the test items that contribute best to the production of score-spread. Items that are answered correctly by only about half the examinees do a great job in spreading out examinees' total-test scores. On the other hand, items answered correctly by a large proportion of examinees, for instance, 80 percent or higher, do not help produce score-spread. Accordingly, the developers of *Alpha*-like standardized achievement tests avoid putting these sorts of high-success items on a test when it's first built, and almost certainly will remove such items when the test is revised. Items that the bulk of students answer correctly do not contribute their share to the production of score-spread.

But here's the catch. Test items on which students perform well will often cover the topics that teachers have emphasized instructionally. The more significant the topic, the more likely it is that teachers will stress the topic. Yet, the better that students perform on items related to any teacher-stressed topics, the less likely it is that those items will be found on the test. There is, therefore, a powerful tendency to remove from traditional standardized achievement tests those items covering the most important things that students learn in school.

Two Extraneous Sources of Score-Spread

If I had a magic wish, it would be that you could spend some time scrutinizing the actual items in today's standardized achievement tests, either national or state-specific tests. What you'd find is an abundance of two types of items that, although they contribute magnificently to the creation of score-spread, have nothing to do with measuring what students learn in school.

The first of these two item-types consists of those items that give an edge to children from middle or upper socioeconomic status (SES) families. To illustrate, in one of the currently used nationally standardized achievement tests there's a sixth-grade science item whose correct answer depends on a student's familiarity with fresh celery. Consideration of the item makes it apparent that sixth-graders whose parents can routinely afford to buy fresh celery will perform better on the item than will sixth-graders whose parents are forced to scrape by on food stamps.

But why, you might reasonably ask, would the people who build standardized achievement tests employ such SES-linked items? The answer is all too simple. SES is a nicely dispersed variable, and a variable that isn't altered overnight. SES-linked items, therefore, typically make a great contribution to the creation of score-spread. But, from an evaluative perspective, SES-linked items measure what students bring to school, not what they learn there.

A second category of item that fails to suitably assess instructional effectiveness consists of those items linked directly to children's *inherited* academic aptitudes such as their verbal, quantitative, or spatial aptitudes. Such items, obviously more suitable for aptitude than achievement tests, assess a student's inborn, genetically determined capacities. To illustrate, there's a fourth-grade mathematics item on a current nationally standardized achievement test that asks students to determine which one of four letters, when folded in half, will have two parts that match exactly. The correct answer for the item is the letter "B." But it should be clear that students who were born with stronger spatial visualization capacities will be better able to cope with such a "mental letter-bending" task. There are also many items on these tests that are clearly dependent on a child's inherited verbal and quantitative aptitudes.

And why, one might ask, would we find these sorts of inheritance-linked items on an achievement test? It's because, as was true with the SES-linked items, they do a great job in producing score-spread. Children's inherited academic aptitudes are nicely distributed. Accordingly, items based on those aptitudes will typically spread out students' scores. But, from an evaluative perspective, inheritance-linked items measure what students bring to school, not what they learn there.

How many of these extraneous items are there in standardized achievement tests? Well, I recently decided to go through a full grade's worth of items from two nationally standardized tests and came up with the following percentages of items that were either SES-linked or inheritance-linked. In reading, I found about 50%; in mathematics, about 20%; in language arts, about 80%; in science, about 80%; and in social studies, about 60%. I really tried to be objective, and I think I was. But even if you were to chop my estimates in half, there were still way too many items that failed to measure what students should learn in school.

Content Mismatches

There's one additional, nontrivial problem with which you should be acquainted, namely, testing-teaching mismatches. The developers of standardized achievement tests have a tough task when trying to representatively sample a set of curricular content (knowledge and skills) in only an hour or so of testing time. Accordingly, there's a high likelihood that the specific content sampled by a standardized achievement test may be seriously inconsistent with local curricular aspirations. A study at Michigan State University, conducted almost two decades ago, suggests that as much as 50% of the items included in a nationally standardized achievement test may cover content that's not even taught in a given locality.

Many educational policymakers make the assumption that a standardized achievement test's content will mesh well with what's supposed to have been taught locally. That assumption is often unwarranted. Teaching-testing mismatches are apt to be more pronounced when national tests are used (because test-makers must cope with considerable national curricular diversity), but teaching-testing mismatches are also

seen with state-specific tests because of the long litanies of content standards the state's curricular experts have chosen.

Mismeasurement-Spawnd Problems

I hope you can see that America's teachers are currently being forced to play a no-win instructional game. They are supposed to boost students' scores on tests that don't measure what children should learn in school. Pressured to raise students' scores on instructionally insensitive tests, is it any wonder that many teachers have (1) reduced their curricular coverage to *only* the content included on a high-stakes test, (2) drilled students *ad nauseum* on items identical or nearly identical to those on the test, or (3) engaged in questionably "relaxed" test-administration practices?

Most teachers can generally do a good instructional job if they have a clear picture of the knowledge and skills that their students should master. Standardized achievement tests, distressingly, fail to provide descriptions of what's actually being measured at a level of clarity sufficient for teachers' instructional planning. And why should such tests? Historically, that's not the purpose of standardized achievement tests. Remember, these descendants of the *Army Alpha* are supposed to provide test-takers' relative standings, not determine how much test-takers learned.

Are the results of standardized achievement tests educationally useful at all? Yes they are. If a standardized achievement test indicates that Chris is relatively strong in mathematics (92nd percentile), but relatively weak in language arts (34th percentile), this is useful information for both teachers and parents. Traditional standardized achievement tests have an important role to play in American education—but that role is *not* to evaluate schools.

Any Alternatives?

Not all standardized achievement tests must be created using an *Alpha*-template. It is possible to build high-stakes standardized achievement tests that not only supply policymakers with credible evidence for accountability purposes, but can also provide teachers with clarified curricular targets that improve teachers' instructional effectiveness.

The nation's school board members, however, must first come to recognize the profound mistake of using traditionally built standardized achievement tests to evaluate a school's instructional success. Measuring school quality with a standardized test is like trying to measure temperature with a tablespoon. It's simply the wrong measurement tool for a very worthwhile task. When educational policymakers understand that they've been using unsuitable tests to evaluate schools, they can then demand the installation of achievement tests more suitable for the evaluation of schools.

What's in a Name?

Most of the misconceptions that today's educational policymakers have regarding standardized achievement tests stems directly from the misleading name that's used to describe those tests. An "achievement test," according to Webster's Dictionary is "a test to measure a person's knowledge or proficiency in something that can be learned or taught." In other words, an educational *achievement* test surely seems to be assessing what students have *learned* in school. But, as I've tried to suggest here, that's not what those tests really do.

They do what they've always done since the *Army Alpha* was born. They sort out examinees. But let's not blame the creators of the *Alpha*. Their test did what it was supposed to do. Instead, let's blame ourselves for letting this increasingly harmful form of educational mismeasurement flourish. It's time to fix it.

References

Popham, W. James. *Testing! Testing! What Every Parent Should Know About School Tests*. Boston: Allyn and Bacon, 2000.

Popham, W. James. *The Truth About Testing: An Educator's Call to Action*. VA: Association for Supervision and Curriculum Development, 2002.