Today’s standardized tests are not the best way to evaluate schools or students

RIGHT TASK

WRONG TOOL

By W. James Popham

Reprinted with permission from American School Board Journal, February 2002 © 2002 National School Boards Association. All rights reserved.
Most Americans, and that includes school board members, believe the best way to evaluate a school is to see how well its students perform on a standardized achievement test. Despite the pervasiveness of this belief, however, it is quite wrong.

Clearly, if education policy makers 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 education policy makers are guided by the wrong measurement data, it is a certainty that misguided policies will follow.

There are reasons why the evaluation of educational quality using standardized achievement tests is improper. If education policy makers begin to recognize the shortcomings of an incorrect evaluative approach, they can then move toward more defensible ways of evaluating schools.

Tests: present and past

Two types of standardized achievement tests will be considered here—namely, national tests and state-specific tests. Currently, five nationally standardized achievement tests are widely used in U.S. public schools. These tests, which include 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 they 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 those 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 focus 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 intelligence test administered to nearly 2 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 United States, 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 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 that has an overriding purpose to provide relative comparisons among test-takers can never be a suitable tool for evaluating what students have learned in school. Here’s why.

For 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.

Let’s suppose a test-developer is faced with a choice between Item A measuring something really important that students should be taught, or 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

The most important factor in evaluating the success of a given school’s staff should surely be how much the 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, 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 are answered correctly by most students 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 teachers will stress it. 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.

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 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 types consists of items that give an edge to children from middle or upper socioeconomic status (SES) families. To illustrate, one of the currently used nationally standardized achievement tests includes 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, a fourth-grade mathematics item on a current nationally standardized achievement test 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. Many items on these tests 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 extraneous items are 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. I found about 50 percent in reading; about 20 percent in mathematics; about 80 percent in language arts and science; and about 60 percent in social studies. I really tried to be objective, and I think I was. But even if you were to chop my estimates in half, way too many items still 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 many as 50 percent of the items included in a nationally standardized achievement test may cover content that's not even taught in a given locality.

Many education policy makers make the assumption that a standardized achievement test's content will mesh well with what's supposed to be 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 these mismatches are also seen with state-specific tests because of the long litigations of content standards the state's curricular experts have chosen.

Mismeasurement 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. 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 do so? 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 im-

### GUIDELINES FOR BUILDING A BETTER TEST

SO, IF TODAY’S TESTS and state-level accountability systems don’t work, what can be done to make them more effective?

A lot, say members of the Commission on Instructionally Supportive Assessment, which has issued a set of guidelines for states to follow in setting up testing programs. The independent commission was convened by five education associations in response to President Bush’s plan to test every child in grades three through eight.

“There is no doubt that the kinds of tests being used in today’s state accountability programs—mandated standardized achievement tests—are causing educational harm, perhaps irreparable harm, to thousands of American children,” said commission chair W. James Popham when the guidelines were released in October.

Popham, a critic of many current state-level accountability programs, said high-stakes testing has “triggered educationally unsound practices.” Today’s programs, he noted, do not measure what students are learning and prevent teachers from focusing instruction.

“Excessive test-focused drill sessions … can surely wipe out the joy that children might otherwise experience from learning,” Popham said.

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 Schools Association are supporting the commission’s findings and recommendations.

Vincent Ferrandino, executive director of NAESP, said tests must “serve a useful purpose to improve instruction, not simply be a high-stakes measure of performance.”

“We don’t want the critical process of test development to be rushed, as it has unfortunately [been] in many states,” Ferrandino said. “We need tests that work for students of all abilities and diversities, and we must give teachers and principals the kind of professional training they need to use the results skillfully.”

NEA President Bob Chase, who leads the nation’s largest teacher union, agreed.

“Tests are an important tool of the trade,” he said. “… But tests are snapshots. They should not lead to snap judgments.”

The commission made nine recommendations:

- **Prioritize a state’s content standards to support effective instruction and assessment.**
- **Describe the content standards clearly and thoroughly so students and teachers know what is needed to demonstrate competence.**
- **Report the state’s assessment of the standards for each student, school, and district.**
- **Provide educators with optional classroom assessment procedures to help measure students’ progress in reaching standards not assessed by state tests.**
- **Monitor the state curriculum to ensure that attention is given to all content standards and subject areas, including those not assessed by state tests.**
- **Provide well-designed assessments that are appropriate for a broad range of students, with accommodations and alternative methods of assessment available when necessary.**
- **Allow test developers a minimum of three years to produce statewide tests.**
- **Ensure that educators receive professional development focused on how to improve children’s learning using instructionally supportive test results.**
- **Find evidence that supports the ongoing improvement of state assessments.**

Each association has posted these recommendations on its Web site, and officials from the five groups promised to “grade” how states are doing in setting up their accountability programs.

If those grades were issued today, said AASA Executive Director Paul Houston, many state legislatures would not pass. Lawmakers are “putting the assessment cart before the instructional horse,” he said. “If educators are accountable for test results, everyone needs to be accountable, including those who make the policy.”

Popham believes that can be done—if states pay attention to the new standards handed down at the federal level.

“It is possible to install accountability tests that will not only provide credible and accurate information about education’s success but can also help improve teaching quality and decision making,” he said. “Better tests will mean the nation’s children can be better taught.”—Glenn Cook, Managing Editor
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 policy makers 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 education policy makers 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 education policy makers have regarding standardized achievement tests stem 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.


CONCERN OVER FEDERAL TESTING

IN 12 YEARS, every student in grades three through eight will be proficient in reading and math.

That’s one promise of the recently reauthorized Elementary and Secondary Education Act. And whether you believe that will happen—or, at least, that schools will make big strides in that direction—depends on how you feel about annual standardized tests.

After months of off-and-on debate, the bill was approved by Congress in mid-December and signed by President Bush, who has backed the tests as a way to “ensure that no child in America is left behind” and that schools are held accountable.

Under the plan, states must test third- through eighth-graders annually using a standardized test chosen by each state. Benchmarks for the state exams will be set by cohorts of students from each state who will also take the National Assessment of Educational Progress.

The law requires states to make steady progress in student achievement, with the goal of 100 percent of the third- through eighth-graders becoming proficient in 12 years. Schools whose scores don’t improve for two consecutive years will receive additional federal money, but if progress continues to lag they could have their staff replaced and curriculum changed.

After one year in a failing school, low-performing students could receive federal funds to pay for transportation to another public school. After two years, public funds could be used to provide private tutoring or summer school.

“These reforms mean new hope for students in failing schools, and new choices for parents who want the best education possible for their children,” said Rep. John Boehner, R-Ohio, who chaired a House-Senate committee that reviewed the bill.

Education groups had mixed reactions to the legislation. The National School Boards Association voiced support for strong assessment programs but said multiple measures were more reliable than a single test. NSBA also said that the roughly $400 million appropriation for testing was insufficient.

“We know that that’s nowhere near what’s needed,” said Reginald Felton, NSBA’s director for Advocacy and Issues Management.

Jill Wynns, board president for the San Francisco Unified School District, said she was skeptical about plans to allow students to transfer out of low-scoring schools, saying her city’s schools were already overcrowded.

And she was wary about the plan to replace the staffs of persistently low-performing schools. San Francisco underwent a massive “reconstitution” program in the 1990s, but almost all the reconstituted schools are on the state’s “accountability list” of troubled schools, Wynns said.

“We’ve already done that,” she said. “It didn’t work.”

—Lawrence Hardy, Associate Editor

Reprinted with permission from American School Board Journal, February 2002 © 2002 National School Boards Association. All rights reserved.