

THE SCORE-BOOSTING GAME: EVERYBODY LOSES!*

W. James Popham
University of California, Los Angeles

“The game’s afoot, Watson,” exclaims Sherlock Holmes, and thereby informs his colleague, Dr. Watson, that one of their crime-solving adventures is underway. Well, there’s currently another game emphatically afoot in American education. It’s called the *score-boosting game*, and it seems to be sweeping the nation.

The aim of the score-boosting game is for teachers to increase their students’ performances on high-stakes tests—typically standardized achievement tests such as the *Stanford Achievement Tests* or the *Iowa Tests of Basic Skills*. Unfortunately, the score-boosting game is one that teachers will almost certainly lose. Even worse, it’s a game destined to make losers out of their students. Indeed, the score-boosting game, as it’s currently being played, is one in which everybody loses. What a bizarre game!

The Game’s Origin

The score-boosting game was born because of a widely held but seriously mistaken belief that students’ scores on important tests accurately reflect educational quality. Schools whose students score well on, for example, the *Comprehensive Tests of Basic Skills* are regarded as effective. Schools whose students score poorly on such tests are regarded as ineffective.

And it doesn’t require a gigantic jump in logic to conclude that the teachers in an effective school must surely be effective, whereas the teachers in an ineffective school must surely be ineffective. Low-performing schools, or so the argument goes, are necessarily staffed by low-performing teachers. Such arguments are brimming with sound logic. But even though logical, they are altogether wrong.

They’re wrong because *inappropriate tests* are being used to arrive at estimates of a school’s effectiveness. Standardized achievement tests, although quite suitable for a number of important educational purposes, do not yield a proper picture of a school staff’s instructional quality. Far too many items on standardized achievement tests fail to measure what students are supposed to learn in school. Rather, those items measure things chiefly influenced by the nature of a school’s student body. To evaluate a school’s staff on the basis of test scores reflecting the composition of that school’s students is so obviously unsound it is almost laugh-inducing. And yet, that’s precisely the rationale for the score-boosting game’s existence.

Educational policymakers, whether legislators or members of state boards of education, really think that children’s scores on standardized achievement tests are chiefly attributable to what’s taught in school. I believe the people behind the score-boosting game are, for the most part, well-intentioned folks who truly want our schools

* A version of this essay appeared in the *American School Board Journal*, Vol. 187, No. 6, June 2000, pp. 36-39.

to be better. But those architects of the score-boosting game made a major mistake when they tried to use students' scores on assessments such as the *Stanford Achievement Tests* to signify educational success. Briefly, let me explain why.

Evaluative Shortcomings of Standardized Achievement Tests

Testing-teaching mismatches. There are three major reasons that standardized achievement tests fail to supply an accurate estimate of a school staff's instructional effectiveness. The first of these is that there are typically substantial mismatches between what's assessed by one of these tests and what's supposed to be taught in a particular school. Given the large bodies of knowledge and the substantial array of cognitive skills that must be taught to students, as a practical matter the builders of standardized achievement tests must *sample* such content. To test *all* of the to-be-taught content would take far too much testing time. Our students would spend so much time taking tests that they'd be receiving social security checks before getting a high school diploma! The effect of such content-sampling is that what's actually measured on the standardized test used in a specific school may be assessing things that weren't even taught in that school.

A landmark study carried out at Michigan State University some years ago suggests that between 50 and 80 percent of what is assessed on a standardized achievement test *is not even taught—or even supposed to be taught*—in many of the schools using that test! Now, if you're a teacher in a school whose effectiveness is being determined by a test in which half or more of the test's items measure things you weren't even supposed to teach, how accurate do you think any score-based judgments about your school's instructional success are likely to be?

Removal of key content. And now for reason number two. Standardized achievement tests are rooted in their World War I origins during which the *Army Alpha* was employed to distinguish among nearly 2,000,000 men with respect to their likelihood of being successful army officers. The comparative assessment strategy used in the *Alpha*, and still employed in today's standardized achievement tests, is dependent on the existence of a substantial *score-spread* among those who take the test. If the scores are sufficiently spread out, then we can tell who scores at the 85th percentile and who scores at the 47th percentile according to the performances of a suitably spread out norm group.

Test items that do the best job in spreading out students' scores are those that are answered correctly by only about half the students. To illustrate, items answered correctly by between 40 and 60 percent of the students do a super job in creating the score-spread so requisite for fine-grained normative comparisons. Items that too many students answer correctly, for instance, items answered correctly by 80 or 90 percent of the examinees, tend to be removed from a standardized achievement test whenever it is revised every half-dozen or so years.

Unfortunately, and here's the snag, test items on which students perform *well* often cover the very content that, because of its importance, teachers *stress*. The better the teacher does in promoting students' mastery of that important content, the more likely it is that students will correctly answer the items covering this teacher-stressed content. And this means there's much less likelihood a test item covering that significant content will be in the next revision of the test. The item is apt to be dumped because too many students answered it correctly. It is not contributing its share to creating score-spread.

Not all important content, of course, will have been jettisoned from an oft-revised standardized achievement test. But there is definitely a technical tendency, *because of the relentless quest for score-spread*, to remove items covering the most important things that teachers teach. That's a powerful reason standardized achievement tests shouldn't be used to evaluate educational quality.

Confounded causality. The third reason that students' scores on standardized achievement tests should not be used to evaluate the quality of schooling also stems from the overriding need of the tests' developers to produce sufficient score-spread. Some items on standardized achievement tests do, in fact, measure the kinds of things teachers ought to be teaching. But some items don't. To be more specific, students' answers to far too many items on a standardized achievement test depend on a student's *socioeconomic status* (SES) or the student's inherited academic aptitudes.

As a consequence, when students' test scores are high, we can't tell whether those scores are due to good teaching, to a good family environment, or to good genes. The same ambiguity exists for students who get low scores. Causality of student performance is seriously confounded because we can't sort out what proportion of a student's test performance is due to teaching, to SES, or to inherited academic aptitudes.

To give you an idea of the kinds of test items that are strongly influenced by a child's SES, consider the sixth-grade science item in Figure 1. This item is only slightly, and unimportantly, modified from an actual item you will find in a currently published nationally standardized achievement test.

Insert Figure 1 about here.

Look carefully at the item seen in Figure 1, then think about a student who comes from an affluent family, a family that can often afford to buy fresh celery at the supermarket and can purchase pumpkins annually for Halloween-carving. That student is far more likely to answer this item correctly than is a child whose family is getting by on food stamps. Students' performances on this item, on average, will be meaningfully linked to their family's SES. The probabilities are quite clear. Kids from advantaged homes will tend to do better on this item than will kids from disadvantaged homes.

There are currently too many SES-linked items such as this on standardized achievement tests. But why, you might ask, is this so? Why would test-developers ever do something so patently unfair? The answer is one you probably won't like all that well. It's because SES, from a test-developer's perspective, is a delightfully spread out variable—and SES isn't rapidly modifiable. SES-linked items will almost certainly produce plenty of score-spread.

In passing, let me point out that many of the language arts items you'll find on most standardized achievement tests are going to be far easier to answer correctly if you're a child who grew up in a family where standard American English was routinely used than if your family used another language or employed non-standard English. Again, especially in language arts, children from affluent English-speaking backgrounds have a gigantic leg up on less fortunate youngsters.

Let's turn now to *inherited academic aptitudes*, and I'm thinking here of the inherited differences we see in different children's spatial, verbal, or quantitative potentials. Now, please consider an illustrative item that, in a significant manner, taps one of these inherited academic aptitudes. Take a look at the item in Figure 2. It's only slightly modified from an actual item that currently appears in the fourth-grade mathematics section of a widely used standardized achievement test.

Insert Figure 2 about here.

To answer correctly the item in Figure 2, a child needs to possess a reasonable spatial visualization capacity. But, although improvable, spatial visualization aptitude is something children are born with. In general, boys are better at coping with this sort of item than girls. Besides, what sensible teacher spends time instructing students on how to engage in "mental letter folding?" It's not a skill that adults are often obliged to use.

There are too many such aptitude-linked items in standardized achievement tests. And again you might ask, why is this so? Why would test-developers use such items? They do so because inherited academic aptitudes such as a child's verbal, quantitative, or spatial capacities are, from a test-developer's perspective, marvelously spread out—and those aptitudes aren't readily modified. Many aptitude-linked items, the kind we formerly found on intelligence tests, are comfortably camouflaged in today's standardized achievement tests. And such items will assuredly produce plenty of score-spread.

You might conclude that these three problems exist only in nationally standardized achievement tests, and could readily be eliminated if a state's educational officials simply created their own customized test to mesh better with the state's curricular preferences. Don't be too optimistic. By and large, state-customized tests are built by the very same companies that created the nationally standardized

achievement tests. And the test-development techniques used (at least from the state-customized tests I've observed) often yield state-tailored tests that function pretty much the way national standardized achievement tests work.

The Score-Boosting Game's Rules

Hopefully, you now recognize today's teachers are being asked to increase students' scores on assessment instruments that may work wonderfully for making *norm-referenced* comparisons among students, but do a dismal job of evaluating instructional quality. Nonetheless, the score-boosting game is still every bit as afoot as it can be. So, wrong tests or not, what are the game's rules?

There are three main rules in the score-boosting game. One is a positive; two are negative. First off, and most important, teachers are supposed to increase students' scores on whatever tests are believed by the relevant policymakers to reflect student learning. More often than not, these improvements in students' scores are supposed to show up from year to year as when, for example, the scores of this year's fifth-graders are contrasted with the scores of last year's fifth-graders. The goal of the game is for teachers to promote a year-to-year rise in students' scores.

But there are two negative game-rules as well. First, in order to promote test-score increases, teachers are prohibited from teaching directly toward the actual items on a test. Nor are teachers supposed to direct their instruction toward items that are too similar to a test's actual items. This rule is illustrated in Figure 3 where you can see that a teacher's instruction *should* be directed toward the body of knowledge and/or skills *represented* by a test, and *should not* be directed toward the test's items themselves.

Insert Figure 3 about here.

A second negative game-rule is that the teacher must not depart from official test-administration precepts, that is, the rules governing the standardized administration of the test. (Classroom teachers, of course, are the test-administrators for almost all of today's standardized achievement tests.) Teachers who allow extra time for students or who provide their students with during-the-test hints about correct answers would obviously be violating this rule.

So, in a nutshell, a teacher who's forced to play the score-boosting game must (Rule 1) *raise students' scores*, (Rule 2) *shun item-teaching*, and (Rule 3) *avoid administration-bending*. The problem is, however, a teacher cannot satisfy Rule 1 without violating Rules 2 or 3 *unless* the test's publishers provide a reasonably clear description of what knowledge and/or skills are *represented* by the test. And that, distressingly, is where the publishers of standardized achievement tests fall down—way down.

I've spent a great deal of time in recent years studying the descriptive information made available to teachers by the publishers of standardized achievement tests. For the most part, such descriptive material is astonishingly deficient *for instructional purposes*. A teacher simply doesn't have an idea about how to focus instruction because there's just not enough clarity about what's being measured.

Yet, the bulk of our nation's educational policymakers really believe that all they need to do is point teachers toward a high-stakes achievement test, then issue a challenge to "Raise scores!" Without sufficiently clear descriptions of the knowledge and skills a test is supposed to represent, such score-raising can't be done—at least it can't be done legitimately.

Playtime

To give you a better idea of the score-boosting perplexities that today's teachers face, I want you to do a little simulated game-playing. Please pretend that you are a sixth-grade teacher. Moreover, imagine you are under substantial pressure from higher-ups to boost your sixth-graders' standardized achievement test scores.

I'll describe a scenario, using an actual example of the information available to sixth-grade teachers when they consult the descriptive materials provided by test publishers. For this simulation, I'll draw only on the descriptive information provided by a nationally standardized achievement test currently in use. I'll also describe the kinds of test items actually found on the test. (To make more apparent what kinds of descriptive material a teacher would be given, I will quote *and* italicize all such publisher-supplied information.)

To set the stage for this fictional game-playing, assume that you've just finished administering your district-approved standardized achievement test for the first time, and you don't have to return it to the principal until the end of the week. (In other words, you have a day or two to look the test over.) You're trying to decide how to deal with the almost constant reminders from your principal and your superintendent that next year's test scores must climb. Imagine that you can refer to the test itself as well as to the test publisher's documents describing what the test is supposed to measure. Typically, such documents are referred to as guides to classroom planning, interpretive guides, or something similar. In most cases, such guides lay out the objectives on which the test's items are supposedly based. By and large, these sorts of descriptive materials supply teachers with the only idea they will ever have (apart from the actual test items) about what's being measured by the test. Putting it another way, the directly quoted and italicized descriptive language I'll soon supply for your make-believe game-playing is all teachers get to let them know what the test is supposed to assess.

For your simulated play, let's dip into mathematics. One of the major nationally standardized achievement tests in sixth-grade mathematics currently contains five items based on the degree to which a student "*demonstrates an understanding of the process of solving conventional and non-routine problems.*" Non-routine problems are, by

definition, routine-free. Imaginative test-item writers, by dodging any hum-drum problems, can devise all sorts of exotic problems for students to solve. As a consequence, it will be nearly impossible for you to know what sorts of non-routine problems are likely to appear on a specific achievement test. Your recognition of what's to be assessed will not be much improved even when the test's descriptive materials attempt to set some content constraints by indicating when students answer these problem-solving items they may need to "*identify missing information*" or to "*solve problems using non-routine strategies*." As our imaginary sixth-grade teacher, you'll still be pretty uncertain about the boundaries of the mathematical skill on which this test's five problem-solving items are based.

It's clearly impossible for you to prepare your students for all possible variations of non-routine problems because the nature of such problems is literally infinite. And, remember, *all* you have to rely on for an understanding of what's to be measured is the italicized descriptive information I've cited here. Accordingly, think of how tempted you might be to focus on the test's five problem-solving items, then teach your next year's students to *clobber* those five particular problems.

If you engaged in this sort of item-focused practice, would your next year's sixth-graders be likely to come up with a correct answer to the actual test's five problem-solving items? Of course, they will. But will your students' mastery of these specific five items help demonstrate they can solve other sorts of non-routine problems? Absolutely not. The more you provide atypical instruction for specific types of problems found on the test, the less valid will be any inference about your students' mastery of the objective on which such test items are based.

Moreover, how likely is it that you will be so dismayed by the dysfunctional rules of the score-boosting game that you might become more permissive when you administer the standardized achievement test to next year's sixth-graders? You may reasonably conclude that the score-boosting game's requirements are so unsound that showing a bit of leniency to your students doesn't seem all that inappropriate.

California provides a perfect illustration of how the score-boosting game can foster the sort of classroom instructional corruption that robs children of curricular content those children ought to be receiving. Such content is often elbowed out because of California's score-boosting hysteria. Important concepts are neglected because they aren't included on the state's high-stake tests.

Moreover, every time a teacher breaks the game's item-teaching or administration-bending rules, students' test scores no longer provide a valid indication of how much students have actually learned. Therefore, instructional decisions relying on invalid score-based inferences about students' achievement levels will almost certainly be faulty.

The chief message I'm trying to transmit is quite straightforward. Because of the absence of sufficiently clear descriptors regarding what content is represented by standardized achievement tests, it is impossible for teachers, or their students, to be

winner in the score-boosting game. If you spend *serious* analytic time with the descriptive information supplied by standardized achievement test publishers about what knowledge and skills their tests *represent*, you'll most certainly find such descriptions insufficient for practical, real-world instructional decision-making.

Changing the Game

So what's to be done about this harmful score-boosting game? Well, I propose a two-pronged strategy that starts off with a big push toward *assessment literacy*. First, educators themselves need to become genuinely literate about assessment. Once educators have acquired sufficient measurement moxie, they then need to help policymakers also acquire enough assessment literacy so those policymakers will refrain from doing what has spawned the score-boosting silliness we see all around us. (I've written a couple of books, cited at the essay's close, that could help educators start traveling down the assessment-literacy trail. There are other such books around, but I am uncontrollably partial.)

Secondly, we need to collect compelling, credible evidence that our students are, in fact, learning good things—and learning them well. The accountability genie has been out of the bottle for a long while now, and just won't be stuffed back in. Citizens, and their elected representatives, aren't going to be satisfied with glib assurances from educators that children are learning well. The public wants, and has the right to receive, hard *evidence* of educational quality.

So, abetted by better insights about what constitutes acceptable evidence of school quality, educators need to gather such evidence and show the world they're doing a good job. The score-boosting game, *if the right sorts of tests are used*, can be a boon to both teachers and students. As it's currently being played—everybody loses.

References

Popham, W. James. (2000). *Modern Educational Measurement: Practical Guidelines for Educational Leaders*. Boston: Allyn and Bacon.

Popham, W. James. (1999). *Classroom Assessment: What Teachers Need to Know*. Boston: Allyn and Bacon.

Figure 1

A sixth-grade standardized achievement test item in science:

Scientists have learned that a plant's fruit always contains seeds.
Which of these is not a fruit?

A. apple

* C. celery

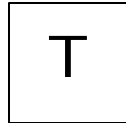
B. pumpkin

D. orange

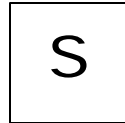
Figure 2

A fourth-grade standardized achievement test item in mathematics:

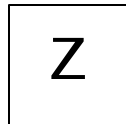
Look at the letters below, then pick the one that can be folded in half so its two parts match exactly.



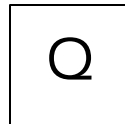
*A



B



C



D

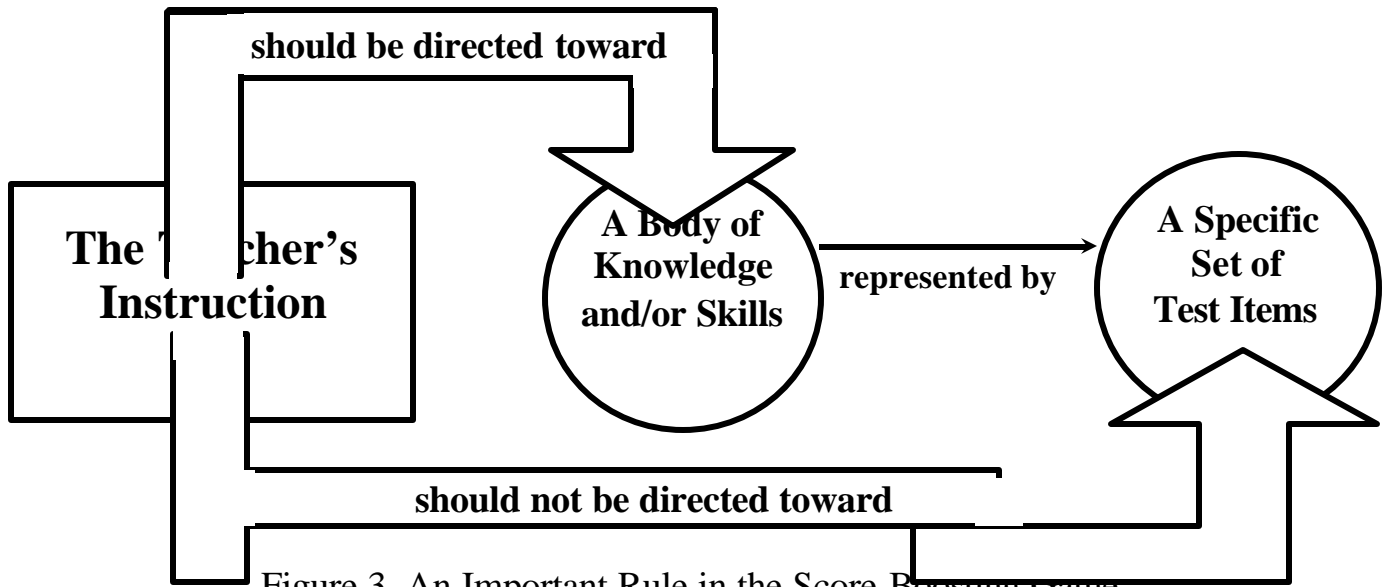


Figure 3. An Important Rule in the Score-Boosting Game.