VEER: JAMES GODMAN

FOR ASSESSMENT

For the last four decades, students' scores on standardized tests have increasingly been regarded as the most meaningful evidence for evaluating U.S. schools. Most Americans, indeed, believe students' standardized-test performances are the only legitimate indicator of a school's instructional effectiveness. Yet, although test-based evaluations of schools seem to occur almost as often as fire drills, in most instances these evaluations are inaccurate. That's because the standardized tests employed are flat-out wrong.

Standardized tests have been used to evaluate America's schools since 1965, when the U.S. Elementary and Secondary Education Act (ESEA) became law. That statute provided for the first major infusion of federal funds into local schools and required educators to produce test-based evidence that ESEA dollars were well spent.

But how, you might ask, could a practice that's been so prevalent for so long be mistaken? Just think back to the many years we forced airline attendants and nonsmokers to suck in secondhand toxins because smoking on airliners was prohibited only during takeoff and landing. Some screwups can linger for a long time.

But mistakes, even ones we've lived with for decades, can often be corrected once they've been identified, and that's what we must do to halt today's wrongheaded school evaluations. If enough educators —and noneducators—realize that there are serious flaws in the way we evaluate our schools, and that those flaws erode educational quality, there's a chance we can stop this absurdity.

By W. James Popham

Instructionally Insensitive

First, some definitions.

A standardized test is any test that's administered, scored, and interpreted in a standard, predetermined manner. Standardized aptitude tests are designed to make predictions about how a test taker will perform in a subsequent setting. For example, the SAT and ACT are used to predict the grades that high school students will earn when they get to college. In contrast, standardized achievement tests indicate how well a test taker has acquired knowledge and mastered certain skills.

Although students' scores on standardized aptitude tests are sometimes unwisely stirred into the school-evaluation stew, scores on standardized achievement tests are typically the ones used to judge a school's success. Two kinds of standardized achievement tests commonly used for school evaluations are ill suited for that measurement.

The first of these categories are nationally standardized achievement tests like the Iowa Tests of Basic Skills, which employ a comparative measurement strategy. The fundamental purpose of all such tests is to compare a student's score with the scores earned by a previous group of test takers (known as the "norm group"). It can then be determined if Johnny scored at the 95th percentile on a given test (attaboy!) or at the 10th percentile (son, we have a problem).

Because of the need for nationally standardized achievement tests to provide fine-grained, percentile-by-percentile comparisons, it is imperative that these tests produce a considerable degree of score-spread—in other words, plenty of differences among test takers' scores. So producing score-spread often preoccupies those who construct standardized achievement tests.

Statistically, a question that creates the most score-spread on standardized achievement tests is one that only about half the students answer correctly. Over the years, developers of standardized achievement tests have learned that if they can link students' success on a question to students' socioeconomic status (SES), then that item is usually answered correctly by about half of the test takers. If an item is answered correctly more often by students at the upper end of the socioeconomic scale than by lower-SES kids, that question will provide plenty of score-spread. After all, SES is a delightfully spread-out variable and one that isn't quickly altered. As a result, in today's nationally standardized achievement tests, there are many SES-linked items.

Unfortunately, this kind of test tends to measure not what students have been taught in school but what they bring to school. That's the reason there's such a strong relationship between a school's standardized-test scores and the economic and social makeup of that school's student body. As a consequence, most nationally standardized achievement tests end up being instructionally insensitive. That is, they're unable to detect improved instruction in a school even when it has definitely taken place. Because of this insensitivity, when students' scores on such tests are used to evaluate a school's instructional performance, that evaluation usually misses the mark.

A second kind of instructionally insensitive test is the sort of standardized achievement test that has been developed for accountability by many states during the past two decades. Such tests were typically created to better assess students' mastery of the officially approved skills and knowledge. Those skills and knowledge, sometimes referred to as goals or curricular aims, are usually known these days as content standards. Thus, such state-developed standardized assessments—like the Florida Comprehensive Assessment Test (FCAT)—are frequently described as "standards-based" tests.

Some tests measure not what students have been taught in school but what they bring to school.

Because these customized standards-based tests were designed (almost always with the assistance of an external test-development contractor) to be aligned with a state's curricular aspirations, it would seem that they would be ideal for appraising a school's quality. Unfortunately, that's not the way it works out. When a state's education officials decide to identify the skills and knowledge that students should master, the typical procedure for doing so hinges on the recommendations of subject-matter specialists from that state. For example, if authorities in Ohio or New Mexico want to identify their state's official content standards for mathematics, then a group of, say, 30 math teachers, math-curriculum consultants, and university math professors are invited to form a statewide content-standards committee. Typically, when these committees attempt to identify the skills and knowledge the students should master, their recommendation—not surprisingly—is that students should master everything. These committees seem bent on identifying skills that they fervently wish students would possess. Regrettably, the resultant litanies of committee-chosen content standards tend to resemble curricular wish lists rather than realistic targets.

Whether or not the targets make sense, there tend to be a lot of them, and the effect is counterproductive. A state's standards-based tests are intended to evaluate schools based on students' test performances, but teachers soon become overwhelmed by too many targets. Educators must guess about which of this multitude of content standards will actually be assessed on a given year's test. Moreover, because there are so many content standards to be assessed and only limited testing time, it is impossible to report any meaningful results about which content standards have and haven't been mastered.

After working with standards-based tests aimed at so many targets, teachers understandably may devote less and less attention to those tests. As a consequence, students' performances on this type of instructionally insensitive test often become dependent upon the very same SES factors that compromise the utility of nationally standardized achievement tests when used for school evaluation.

Wrong Tests, Wrong Consequences

Bad things happen when schools are evaluated using either of these two types of instructionally insensitive tests. This is particularly true when the importance of a school evaluation is substantial, as it is now. All of the nation's public schools are evaluated annually under the provisions of the federal No Child Left Behind Act (NCLB). Not only are the results of the NCLB school-by-school evaluations widely disseminated, there are

also penalties for schools that receive NCLB funds yet fail to make sufficient test-based progress. These schools are placed on an improvement track that can soon "improve" them into nonexistence. Educators in America's public schools obviously are under tremendous pressure to improve their students' scores on whatever NCLB tests their state has chosen.

With few exceptions, however, the assessments that states have chosen to implement because of NCLB are either nationally standardized achievement tests or state-developed standards-based tests—both of which are flawed. Here, then, are three adverse classroom consequences seen in states where instructionally insensitive NCLB tests are used:

• Curricular reductionism.
In an effort to boost their students' NCLB test scores, many teachers jettison curricular content that—albeit important—is not apt to be covered on an upcoming test. As a result, students end up educationally shortchanged.

• Excessive drilling.
Because it is essentially impossible to raise students' scores on instructionally insensitive tests, many teachers—in desperation—require seemingly endless practice with items similar to those on an approaching accountability test. This dreary drilling often stamps out any genuine joy students might (and should) experience while they learn.

• Modeled dishonesty.
Some teachers, frustrated by being asked to raise scores on tests deliberately designed to preclude such score raising, may be tempted to adopt unethical practices during the administration or scoring of accountability tests. Students learn that whenever the stakes are high enough, the teacher thinks it's OK to cheat. This is a lesson that should never be taught.

These three negative consequences of using instructionally insensitive standardized tests as measuring tools, taken together, make it clear that today's widespread method of judging schools does more than lead to invalid evaluations. Beyond that, such tests can dramatically lower the quality of education.

An Antidote

Is it possible to build accountability tests that both supply accurate evidence of school quality and promote instructional improvement? The answer is an emphatic yes. In 2001, prior to the enactment of NCLB, an independent national study group, the Commission on Instructionally Supportive Assessment, identified three attributes that an "instructionally supportive" accountability test must possess:

• A modest number of supersignificant curricular aims.
To avoid overwhelming teachers and students with daunting lists of curricular targets, an instructionally supportive accountability test should measure students' mastery of only an intellectually manageable number of curricular aims, more like a half-dozen than the 50 or so that a teacher may encounter today. However, because fewer curricular benchmarks are to be measured, they must be truly significant.

• Lucid descriptions of aims.
An instructionally helpful test must be accompanied by clear, concise, and teacher-palatable descriptions of each curricular aim to be assessed. With clear descriptions, teachers can direct their instruction toward promoting students' mastery of skills and knowledge rather than toward getting students to come up with correct answers to particular test items.

• Instructionally useful reports.
Because an accountability test that supports teaching is focused on only a very limited number of challenging curricular aims, a student's mastery of each subject can be meaningfully measured, letting teachers determine how effective their instruction has been. Students and their parents can also benefit from such informative reports.

These three features can produce an instructionally supportive accountability test that will accurately evaluate schools *and* improve instruction. The challenge before us, clearly, is how to replace today's instructionally insensitive accountability tests with better ones. Fortunately, at least one state, Wyoming, is now creating its own instructionally supportive NCLB tests. More states should do so.

What You Can Do

If you want to be part of the solution to this situation, it's imperative to learn all you can about educational testing. Then learn some more. For all its importance, educational testing really isn't particularly complicated, because its fundamentals consist of commonsense ideas, not numerical obscurities. You'll not only understand better what's going on in the current mismeasurement of school quality, you'll also be able to explain it to others. And those "others," ideally, will be school board members, legislators, and concerned citizens who might, in turn, make a difference. Simply hop on the Internet or head to your local library and hunt down an introductory book or two about educational assessment. (I've written several such books that, though not as engaging as a crackling good spy thriller, really aren't intimidating.)

With a better understanding of why it is so inane—and destructive—to evaluate schools using students' scores on the wrong species of standardized tests, you can persuade anyone who'll listen that policy makers need to make better choices. Our 40-year saga of unsound school evaluation needs to end. Now.

This dreary drilling often stamps out any genuine joy students might (and should) experience when they learn.

W. James Popham, *who began his career in education as a high school teacher in Oregon, is professor emeritus at the University of California–Los Angeles School of Education and Information Studies. Author of 25 books, he is a former president of the American Educational Research Association. Write to letters@edutopia.org.*

HOT LINK

Take the next step toward a better understanding of assessment by visiting the Edutopia Web site, where you'll find articles and documentaries on alternative forms of assessment, interviews and opinion pieces by experts in the field, and a wealth of useful and informative resources, including an instructional module on building an evidence-based assessment.
• www.edutopia.org/assessment