Measuring Up

What Educational Testing Really Tells Us

ILLUSTRATED BY PAUL ZWOLAK

BY DANIEL KORETZ

Educational testing is ubiquitous in America, and its importance is hard to overstate. Tests have a powerful influence on public debate about many social concerns, such as economic competitiveness, immigration, and racial and ethnic inequalities. And achievement testing seems reassuringly straightforward and commonsensical: we give students tasks to perform, see how they do on them, and thereby judge how successful they or their schools are.

This apparent simplicity, however, is misleading.

Test scores do not provide a direct and complete measure of educational achievement. Rather, they are incomplete measures, proxies for the more comprehensive measures that we would ideally use, but that are almost always unavailable to us. There are two reasons for the incompleteness of achievement tests. The first, which has been stressed by careful developers of standardized tests for more than half a century, is that these tests can measure only a subset of the goals of education. Some goals, such as the motivation to learn, the inclination to apply school learning to real situations, the ability to work in groups, and some kinds of complex problem solving, are not very amenable to large-scale standardized testing. Others can be tested, but are not considered a high enough priority to invest the time and resources required. The second reason for the incompleteness of achievement tests—and the one that I will focus on here—is that even in assessing the goals that we decide to measure and that can be measured well, tests are generally very small samples of behavior that we use to make estimates of students' mastery of very large domains of knowledge and skill.

The accuracy of these estimates depends on several factors, one of the most important being careful sampling of content and skills. For example, if we want to measure the mathematics proficiency of eighth graders, we need to specify what knowledge and skills we mean by "eighth-grade mathematics." We might decide that this subsumes skills in arithmetic, measurement, plane geometry, basic algebra, and data analysis and statistics, but then we would have to decide which *aspects* of algebra and plane geometry matter and how much weight

should be given to each component (e.g., do students need to know the quadratic formula?). Eventually, we end up with a detailed map of what the test should include, often called "test specifications" or a "test blueprint," and the developer writes test items that sample from it.

But that is just the beginning. The accuracy of a test score depends on a host of often arcane details about the wording of items, the wording of "distractors" (wrong answers to multiple-choice items), the difficulty of the items, the rubric (criteria and rules) used to score students' work, and so on. The accuracy of a test score also depends on the attitudes of the test takers—for example, their motivation to perform well. It also depends, as we shall see later, on how schools prepare students for the test. If there are problems with any of these aspects of testing, the results will provide misleading estimates of students' mastery of the larger domain.

A failure to grasp this fact is at the root of widespread misunderstandings—and misuses—of test scores. It has often led policymakers astray in their efforts to design productive testing and accountability systems. By placing too much emphasis on test scores, they have encouraged schools to focus instruction on the small sample actually tested rather than the broader set of skills the mastery of which the test is supposed to signal.

To make the principles of testing concrete, let's construct a hypothetical test. Suppose that you publish a magazine and have decided to hire a few college students as interns to help out. You receive a large number of applicants and have decided that one basis for selecting from among them is the strength of their vocabularies. How do you determine that? Conversations with them will help, but may not be sufficient because they are not uniform: a conversation with one applicant may afford more opportunities for using advanced vocabulary than a conversation with a second one. So you decide to construct a standardized test of vocabulary.* You would then confront a serious difficulty: although many teachers and parents may find this fact remarkable in the light of their own experience, the typical adolescent has a huge working vocabulary. Clearly, you will have to select a sample of words to put into your test. In practice, you can get a reasonably good estimate of the relative strengths of applicants' vocabularies by testing them on a small sample of words, if those words are chosen carefully. Assume you will use 40 words, which would not be an unusual number in an actual vocabulary test.

The box below gives the first few words from three lists that you could use to select words for your test.

A	B	C
siliculose	bath	feckless
vilipend	travel	disparage
epimysium	carpet	minuscule

Which list would you use? Clearly not list A, which comprises specialized, very rarely used words. Everyone would receive a score of zero or nearly zero, and that would make the test useless: you would gain no useful information about the relative strengths of their vocabularies. List B is no better. Everyone would obtain a perfect or nearly perfect score. Therefore you would construct your test from list C, which comprises words that some applicants would know and others not.

In this example, the fact that a test is merely a sample of a larger domain is clear. But is sampling always as serious a problem as it is in this contrived example? For the most part, yes.† The tests that are of interest to policymakers, the press, and the public at large entail substantial sampling because they are designed to measure sizable domains, ranging from knowledge acquired over a year of study in a subject to cumulative mastery of material studied over several years.

Returning to the vocabulary test: what would have happened if you had chosen words differently, while keeping them at the same level of difficulty? To make this concrete, assume that you selected all three of the words shown in list C, and that I was also constructing a vocabulary test, but I dropped *feckless* and used *parsimonious* instead. For the sake of discussion, assume that these two words are equally difficult.

What would be the impact of administering my test rather than yours? Over a large enough number of applicants, the average score would not be affected at all, because the two words in question are equally difficult. However, the scores of some individual students would be affected. Even among students with comparable vocabularies, some would know *feckless* but not *parsimonious*, and vice versa.

This illustrates one source of *measurement error*, which refers to inconsistency in scores from one measurement to the next. To some degree, the ranking of your applicants will depend on which words you select from list C, and if you tested applicants repeatedly using different versions of your test, the rankings would vary a little. Another source of measurement error is the fluctuation over time that would occur even if the items were the same. Students have good and bad days. For example, a student might sleep well before one test date but be too anxious to sleep well another time. Or the examination room may be overheated one time but not the next. Yet another source of measurement error is inconsistencies in the scoring of students' responses.

Obviously, it's important to try to keep measurement error to a minimum—and that's why test developers are so concerned with *reliability*. Reliable scores show little inconsistency from one measurement to the next—that is, they contain relatively little measurement error. Reliability is often incorrectly used to mean "accurate" or "valid," but it properly refers only to the consistency of measurement. A measure, including a test, can be reliable but inaccurate—such as a scale that consistently reads too high.

So when all is said and done, how justified would you be in drawing conclusions about vocabulary from the small sample of words on your test? This is the question of *validity*, which is the single most important criterion for evaluating achievement testing. In public debate, and sometimes in statutes and regulations as well, we find reference to "valid tests," but tests themselves are not valid or invalid. Rather, inferences based on test scores are valid or not. A given test might provide good support for one inference, but weak support for another. For example, a well-designed end-of-course exam in statistics might provide good support for inferences about students'

* People incorrectly use the term *standardized test*—often with opprobrium—to mean all sorts of things: multiple-choice tests, tests designed by commercial firms, and so on. In fact, it means only that the test is uniform: that is, that all examinees face the same tasks, administered in the same manner, and scored in the same way. The motivation for standardization is to avoid irrelevant factors that might distort comparisons among individuals.

† There are tests that are not samples of a larger domain. For example, a teacher may want to know whether her class has mastered the list of vocabulary words presented in the past week. She would not be trying to draw any conclusions about students' overall vocabularies, and she would be happy indeed if most students got most of the words right.

mastery of basic statistics, but very weak support for conclusions about mastery of mathematics more broadly. The question to ask is: how *well supported* is the conclusion?

None of the preceding is particularly controversial. These fundamentals of testing may not be well known outside the testing community, but inside that community they are widely agreed upon. The next and final step in this hypothetical exercise, however, is contentious indeed.

Suppose you are kind enough to share with me your test of 40 words. And suppose I intercept every single applicant en route to taking your test, and I give each one a short lesson on the meaning of every word on your test. What would happen to the validity of inferences you might want to base on your test scores?

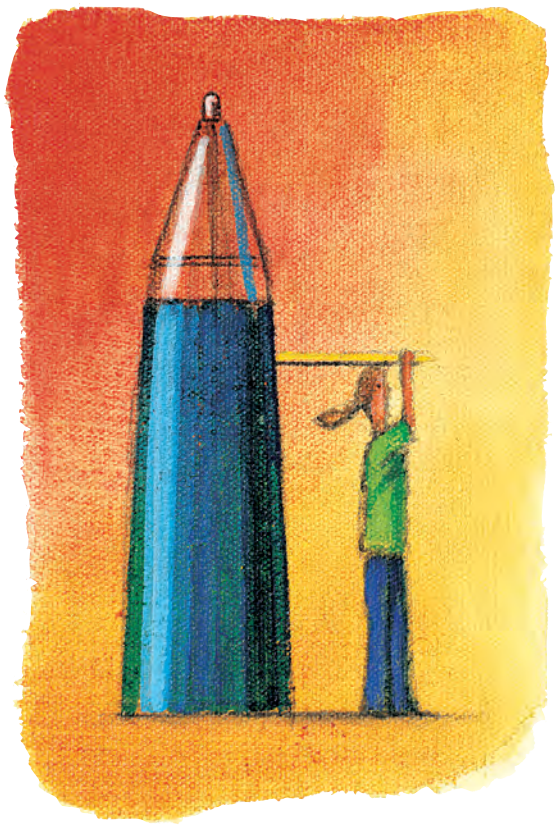
Clearly, your conclusions about which applicants have stronger vocabularies would now be wrong. Most students would get high scores, regardless of their actual vocabularies. Students who paid attention during my mini-lesson would outscore those who did not, even if their actual vocabularies were weaker. Mastery of the small sample of 40 words would no longer represent variations in the students' actual working vocabularies.

This last step—teaching the specific content of the test, or material close enough to it to undermine the representativeness of the test—illustrates the contentious issue of *score inflation*, which refers to increases in scores that do not signal a commensurate increase in proficiency in the domain of interest. Inflation of scores in this case did not require any flaw in the test, and it did not require that the test focus on unimportant material. The 40 words were fine. My response to those 40 words—my form of test preparation—was not.

In real-world testing programs, issues of score inflation and test preparation are far more complex than this example suggests. So let's set aside our vocabulary test and take a closer look at what I believe should be a very serious concern among educators and policymakers: how to prepare for tests.

Test preparation has been the focus of intense argument for many years, and all sorts of different terms (like "teaching the test" and "teaching to the test") have been used to describe both good and bad forms. I think it's best to ignore all of this and to distinguish instead between seven different types of test preparation: (1) working more effectively, (2) teaching more, (3) working harder, (4) reallocation, (5) alignment, (6) coaching students, and (7) cheating.

The first three are what some proponents of high-stakes testing want to see. Clearly, if educators find ways to work more effectively—for example, developing better curricula or teaching methods—students are likely to learn more. Up to a point, if teachers spend more time teaching, achievement is likely to rise. The same is true of working harder in school, although this can be carried too far. For example, it is not clear that depriving young children of recess, which some schools are now doing in an effort to raise scores, is effective, and in my opinion it is undesirable regardless. Similarly, if students' workload becomes excessive, it may interfere with learning and may also generate an aversion to learning. But if not carried to excess, these three forms of test preparation can be expected to produce real gains in achievement that would appear not only in the test scores used for accountability, but on other tests and outside of school as well.



At the other extreme, cheating is unambiguously bad. But what about reallocation, alignment, and coaching? All three can produce real gains, score inflation, or both. Reallocation refers to shifting instructional resources—classroom time, homework, parental nagging, whatever—to better match the content of a specific test. A quarter century of studies confirm that many teachers reallocate instruction in response to tests. And some studies have found that school administrators reassign teachers to place the most effective ones in the grades in which important tests are given.[1]

Is reallocation good or bad? Does it generate real gains in achievement or score inflation? This depends on what gets more emphasis, *and what gets less*. Some reallocation is desirable and is one of the goals of testing programs. For example, if a ninth-grade math test shows that students do relatively poorly in solving basic algebraic equations, one would want their teachers to put more emphasis on such equations. The rub is that devoting more resources to topic A entails fewer resources for topic B.

Scores become inflated when topic B—the material that gets less emphasis as a result of reallocation—is also an important part of the domain. If teachers respond to a test by de-emphasizing material that is important to the domain but is not given much weight on the particular test, scores will become inflated. Performance will be weaker when students take another test that places emphasis on those parts of the domain that have been neglected.

Alignment is a lynchpin of policy in this era of standards-based testing. Tests should be aligned with standards, and instruction should be aligned with both. And alignment is seen by many as insurance against score inflation, but this is incorrect. Alignment is just reallocation by another name. Whether alignment inflates scores also depends on the importance of the material that is de-emphasized. And research has shown that standards-based tests are not immune to this problem. These tests are still limited samples from larger domains, and therefore focusing too narrowly on the content of the specific test can inflate scores.

Coaching students refers to focusing instruction on small details of the test, many of which have no substantive meaning. Coaching need not inflate scores. If the format or content of a test is sufficiently unfamiliar, a modest amount of coaching may even increase the validity

of scores. For example, the first time young students are given a test that requires filling in bubbles on an answer sheet that is going to be scored by a machine, it is worth spending a very short time familiarizing them with this procedure before they start the test.

Most often, however, coaching students either wastes time or inflates scores. A good example is training students to use a process of elimination in answering multiple-choice questions. A *Princeton Review* test-prep manual urges students to do this because "it's often easier to identify the wrong answers than to find the *correct* one."[2] What's wrong with this? The performance gains generated depend entirely on using multiple-choice items. Of course, when students need to apply their knowledge in the real world outside of school, the tasks are unlikely to appear in the form of a multiple-choice item.

This example shows that inflation from coaching is in one respect unlike inflation from reallocation. Reallocation inflates scores by making performance on the test unrepresentative of the larger domain, but it does not distort performance on the material tested. (If I taught applicants the vocabulary words on your test, they would know those words—but their scores on the test would not be good estimates of their overall vocabulary knowledge.) In contrast, coaching can exaggerate performance on the tested material. In the example just given, students who are taught to use the process of elimination as a method for "solving" certain types of equations will know less about those types of equations than their performance on the test indicates.

So what distinguishes good and bad test prep? The acid test is whether the gains in scores produced by test preparation truly represent meaningful gains in student achievement. We should not care very much about a score on a particular test. What we should be concerned about is the knowledge and skills that the test score is intended to represent. Gains that are specific to a particular test and that do not generalize to other measures of the domain and to performance in the real world are worthless.

* * *

This brings me to a final, and politically unpalatable, piece of advice: we need to be more realistic about using tests as a part of educational accountability systems. Systems that simply pressure teachers to raise scores on one test (or one set of tests in a few subjects) are not likely to work as advertised, particularly if the increases demanded are large and inexorable. They are likely instead to produce substantial inflation of scores and a variety of undesirable changes in instruction, such as excessive focus on old tests, inappropriate narrowing of instruction, and a reliance on test-taking tricks.

I strongly support the goal of improved accountability in public education. I saw the need for it when I was an elementary school and junior high teacher, many years ago. I saw it as the parent of two children in school. Nothing in more than a quarter century of education research has led me to change my mind on this point. And it seems clear that student achievement must be one of the most important things for which educators and school systems should be accountable. However, we need an effective system of accountability, one that maximizes real gains and minimizes bogus gains and other negative side effects. Even a very good achievement test will leave many aspects of school quality unmeasured. Some hard-core advocates of high-stakes testing disparage this argument as "anti-testing," but it is a simple statement of fact, one that has been recognized within the testing profession for generations.

So how should you use scores to help you evaluate a school? Start by reminding yourself that scores describe some of what students can do, but they don't describe all they can do, and they don't explain why they can or cannot do it. Use scores as a starting point, and look for other evidence of school quality—ideally not just other aspects of student achievement but also the quality of instruction and other activities within the school. And go look for yourself. If students score well on math tests but appear bored to tears in math class, take their high scores with a grain of salt, because an aversion to mathematics will cost them later in life, even if their eighth-grade scores are good.

Sensible and productive uses of tests and test scores rest on a single principle: don't treat "her score on the test" as a synonym for "what she has learned." A test score is just one indicator of what a student has learned—an exceptionally useful one in many ways, but nonetheless one that is unavoidably incomplete and somewhat error-prone. ❐

Endnotes

1. For a good overview of some of the most important research on teachers' and principals' responses to testing, see Brian M. Stecher, "Consequences of Large-Scale, High-Stakes Testing on School and Classroom Practice," in *Making Sense of Test-Based Accountability in Education*, ed. Laura S. Hamilton, Brian M. Stecher, and Stephen P. Klein (Santa Monica, CA: Rand, 2002), **http://www.rand.org/pubs/monograph_reports/MR1554**.

2. Jeff Rubenstein, *Princeton Review: Cracking the MCAS Grade 10 Math* (New York: Random House, 2000), 15.

This article, which originally appeared in the Fall 2008 issue of *American Educator*, was adapted from Daniel Koretz's book, *Measuring Up: What Educational Testing Really Tells Us*. Detailed but nontechnical, the book addresses the common misunderstandings and misuses of standardized tests, and offers sound advice for using tests responsibly. To learn more, go to **www.hup.harvard.edu/catalog/KORMAK.html**.