11-5-1997

From Multiple Choice to Multiple Choices

Jonathan A. Supovitz
University of Pennsylvania, JONS@GSE.UPENN.EDU

From Multiple Choice to Multiple Choices

Abstract
Are standardized tests an equitable way to measure the achievement of America's children? A fresh, four-year study by the Educational Testing Service of the gender gap on standardized tests concludes that differences in performance between boys and girls are real, but not large, and cut both ways. ("ETS Disputes Charges of Gender Bias," May 14, 1997.) Still, critics of standardized testing, like the National Center for Fair & Open Testing, blast the ETS study as "a smoke screen designed to divert attention from the ongoing problems with the exams they publish."

Disciplines
Disability and Equity in Education | Education | Educational Assessment, Evaluation, and Research | Educational Methods

Comments
This is an online article only and can be found on Education Week.

Published Online: November 5, 1997

Commentary

From Multiple Choice To Multiple Choices

Are standardized tests an equitable way to measure the achievement of America's children? A fresh, four-year study by the Educational Testing Service of the gender gap on standardized tests concludes that differences in performance between boys and girls are real, but not large, and cut both ways. ("ETS Disputes Charges of Gender Bias," May 14, 1997.) Still, critics of standardized testing, like the National Center for Fair & Open Testing, blast the ETS study as "a smoke screen designed to divert attention from the ongoing problems with the exams they publish."

Volumes are printed each year to add to the mountain of data attempting to address the question of test bias. Whole careers are devoted to it. And the newspaper headlines are familiar reading to all of us: "White Students Outperform Minority Students on State Assessment," "Girls Score Lower Than Boys on Districtwide Test." My point is not to belittle these studies, because it is important to hold test publishers' collective feet to the fire. But in an important way, such analyses miss the larger point about bias in today's assessments.

Of course standardized tests are biased. But it is not just standardized tests--any single testing method is biased because it applies just one approach to getting at student knowledge and achievement. Any single testing method has its own particular set of blinders. Since the bias in testing is intrinsic in the form of assessment used, we cannot eliminate this problem simply by changing the questions asked. Rather, we must ask the questions in many different ways. It is time for policymakers and administrators to recognize that even removing all inklings of bias from standardized tests will not remove the bias from today's testing.

This is the conclusion I have reached after examining the relative equity of different assessment types used with primary-grade urban school children. My colleague Robert T. Brennan of Harvard University and I compared the standardized-test performance and portfolio-assessment scores of more than 5,000 1st and 2nd grade students in Rochester, N.Y. That city's 1st graders take the California Achievement Test (CAT-5, level 11), a comprehensive test of reading and writing, including vocabulary, comprehension, word analysis, language mechanics, and language expression. Rochester 2nd graders take the Degrees of Reading Power, or DRP, test, a criterion-referenced test that focuses on student reading and reading-comprehension skills. Both assessments are given in the spring of each year. In addition, all primary-grade students, with guidance from their teachers, amass a language-arts portfolio throughout the school year in which specified samples of student work in reading, writing, and speaking and listening are collected. Teachers then assess the work and assign each student to a developmental stage in reading, writing, and speaking and listening. These stage assignments become a student's portfolio score and are also the basis for report card scores.

Using a sophisticated statistical method called hierarchical linear modeling, which allowed us to directly compare the assessments after adjusting for differences in their reliability, Mr. Brennan and I modeled the relationship between student performance on both the standardized tests and the portfolio, and several student background characteristics that we call "equity characteristics." These include student gender, race or ethnicity, socioeconomic status (as measured by receipt of lunch assistance), and student English-language-learner status (commonly called limited English proficiency). We hypothesized that in the ideal world, knowing a student's equity characteristics would contribute nothing to predicting that student's test performance. In other words, in an unbiased situation, there would be no statistical relationship between students' equity characteristics and test performance. The more equitable the form of assessment, the smaller the amount of variation in test performance would be explained by the equity characteristics.

Our findings are provocative. The equity predictors, taken together, explained approximately 10 percent of the variation in test performance. In the 1st grade, the equity predictors explained less of the variation in test performance on the portfolios than they did on the standardized tests, while in the 2nd grade the reverse was true.

It is time for policymakers and administrators to recognize that even removing all inklings of bias from standardized tests will not remove the bias from testing.

A still more intriguing story emerged when we decomposed the equity predictors. At both grade levels, students' race or ethnicity explained significantly less of the variation in portfolio performance than this characteristic did of students' standardized-test performance. On the other hand, in both the 1st and 2nd grades, a student's gender explained significantly more of the variation in portfolio performance than it did of standardized-test performance, with girls performing significantly better than boys. There were no statistically significant differences in the amount of variation on the two assessment types that was predicted by either our socioeconomic-status variable or English-language-learner status. In sum, these results indicate that portfolios are more biased in terms of gender, while standardized tests are more biased in terms of race or ethnicity.

How do we explain this finding? One hypothesis is that the portfolios simply provide different opportunities for students to demonstrate their language arts skills than the standardized tests do. Thus, in this case, the portfolios gave more opportunities to minority students to demonstrate their knowledge, while the boys were more comfortable with the standardized-test form of assessment than were their female peers. Different assessment forms stress different cognitive abilities and skill experience.

Several other studies, with data from different grade levels and different subjects, have arrived at a similar conclusion. For example, in a 1991 study of 4th grade, hands-on science assessments by Stanford University researchers Richard Shavelson and Gail Baxter, students conducted several scientific inquiries with lab equipment and materials. They also completed corresponding notebooks, computer simulations, paper-and-pencil measures, and a traditional multiple-choice science-achievement test. The researchers found a low correlation between the performance

assessment and multiple-choice test and determined that "for individual students, measures of science achievement are highly sensitive not only to the investigation used, but also to the method used to measure performance."

The effects of a narrow use of assessment go beyond simply a restricted view of children's capabilities. As is widely acknowledged, testing drives instruction. Teachers change their curriculum to prepare students for state and district assessments. So the indicators of student achievement chosen by state- and district-level policymakers send signals to teachers about what they should spend class time preparing their students for. I have observed many classes where students are practicing multiple-choice skills. It is no secret that what gets tested is invariably what gets taught.

The reason for this is that today's large-scale, largely multiple-choice assessments exist in a vacuum. They stand alone, inflating their importance. Since there are no other forms of assessment that, in combinations with standardized tests, can provide a more robust image of a student's capabilities, we have come to rely on one particular type of assessment as *the* measure of student achievement. Standardized tests are the only game in town.

But what other profession looks at just one indicator in making high-stakes judgments? Don't doctors use multiple tests when diagnosing patients? Judges and juries don't rely on one piece of evidence--they scrutinize multiple factors before reaching their verdicts. In my profession, evaluation, we don't rely on just one method to assay effectiveness, we employ multiple methods. So why should judging student achievement be any different? Doesn't it make more sense to find an appropriate balance of different kinds of assessments--open-ended assessments, performance assessments, multiple-choice tests, even portfolios--so that students have different opportunities to show what they know?

The reason test bias is a question that will never be put to rest is that we have no criteria for knowing what a student's true performance is. If we did, we could measure the relative difference away from this "true" benchmark as measured by a specific test. Without this absolute measure, we have no anchors against which to measure performance.

Many educational reforms look toward alternative forms of assessment to give students from different backgrounds the opportunity to "show what they know." Indeed, our study demonstrates this point nicely. Rochester's portfolios expanded the inequities between boys and girls, but reduced the gaps between ethnic groups. Gender and ethnic differences in test performance are probably due less to bias in the items of these tests than to biases inherent in the type of assessment. Based on this, wouldn't an assessment system that included multiple methods of assessment be less biased than any one method alone?

Though multiple-choice tests are the most efficient testing measure yet developed, they, like any single form of assessment used alone, remain limited. What we need are more experiments employing combinations of assessment approaches to arrive at an appropriate melding of test forms both economically feasible and robust enough to minimize the bias inherent in any single measure alone. And alternative forms of assessment must be put through the same rigorous piloting and reliability and validity analyses that multiple-choice tests have undergone.

Shouldn't we find a balance of different kinds of assessments, so that students have different opportunities to show what they know?

In the end, the larger, more intractable sources of disparities in student performance stem from broad social and educational inequities. But within the realm of assessment, the challenge for educators and policymakers is to find the appropriate balance of a variety of assessment forms, so that students of different genders, from different backgrounds, and with different affinities can demonstrate their capabilities.

Educators today must go beyond the minimization of bias in constrained forms of testing. To seek greater equity, we have to develop a plethora of rigorously constructed assessment forms, understanding that each will provide some advantage to certain kinds of students, but that, taken together, they will be a fairer measure of that complex thing we call knowledge. A diverse society deserves a more diverse assessment system.

Jonathan A. Supovitz is a senior researcher at the Consortium for Policy Research in Education at the University of Pennsylvania in Philadelphia. He is the co-author of "Mirror, Mirror on the Wall, Which is the Fairest Test of All," published in the fall issue of the *Harvard Educational Review*, on which this essay is based.