

The High Stakes Of Standardized Testing

Edward Miller

I recently participated in a study, conducted by the National Research Council, of the appropriate uses of standardized tests for making decisions about individual students. Its findings may be of interest to readers who are concerned about the ways in which the technology of testing has become one of the most powerful influences in our education system.

The study committee was charged by Congress with examining the use of test scores for so-called high-stakes purposes, defined as making decisions about tracking, promotion, and graduation. Such uses are proliferating all over the country, and are widely considered an effective tool for whipping the public schools into shape. For example, students in Chicago must now get at least a certain score on the Iowa Test of Basic Skills to be promoted to the next grade. Starting next year, high school students in New York will have to pass the state Regents exam (formerly optional) to get a diploma. The committee found that, while testing can and often does yield valuable information about students' achievement, the nature and limitations of that information are widely misunderstood. Test results, the study concluded, are often used improperly. In the case of high-stakes tests, the effects on individual students' lives may be disastrous.

The committee adopted three basic criteria for determining whether a particular test use is appropriate:

Measurement validity—whether a test is valid for a particular purpose, and whether it accurately measures the test takers' knowledge.

Attribution of cause—whether a student's performance on a test reflects knowledge and skill based on appropriate instruction or is attributable to poor instruction or to such factors as language barriers or disabilities unrelated to the skills being tested.

Effectiveness of treatment—whether test scores lead to placements and other consequences that are educationally beneficial.

These criteria, which were derived from the established standards of the testing profession, reflect a fundamental truth about tests that is well known by experts but generally obscured in public policy debates and news reports: test scores are subject to all kinds of statistical and human error and are therefore very often wrong. Moreover, there is a remarkable lack of agreement in many cases about whether a particular test even measures what it is supposed to measure. But because educational test results are given in numerical form they create a powerful impression of scientific precision—that they are like a thermometer or your blood pressure reading. They are not. They provide only one perspective—and often a very narrow and clouded one—on a student's actual knowledge. This appearance of precision in test scores has been used in many instances to rationalize discriminatory and unfair practices.

The nature of standardized testing, and its history of misuses, leads inexorably to certain conclusions. One is that any use of a test score to justify an educational decision that is likely to harm rather than help the child is, by definition, insupportable. With regard to tracking and promotion, this logic led the study committee to some surprising findings.

After thoroughly examining the research literature on tracking, the group concluded that “students assigned to lowtrack classes are worse off than they would be in other placements.

This form of tracking should be eliminated. Neither test scores nor other information should be used to place students in such classes.”

The committee was similarly troubled by the evidence on “retention”-the practice of making kids repeat a grade. In spite of the popularity of President Clinton’s call to “end social promotion,” the committee found that “grade retention is pervasive in American schools” and that it is usually not educationally beneficial, but leads to lower achievement and higher risk of dropping out. It called for early identification of and remedial programs for students in difficulty as an alternative to holding them back, and it condemned the growing practice of using the results of a single test to determine whether a child should go on to the next grade.

Indeed, the committee concluded that high-stakes decisions of any kind “should not automatically be made on the basis of a single test score.” Other important conclusions were that the use of high-stakes tests to “lead” curricular reform - that is, to get schools to change what and how they teach-tends to corrupt and invalidate the tests, and is fundamentally unfair to students; that large-scale standardized tests should not be used at all in making high-stakes decisions about students below grade three; and that the existing mechanisms for enforcing standards of appropriate test use are inadequate.

The implications of these findings are sobering in light of the growing enthusiasm for more testing as the answer to the intractable problems of school reform in the US. The parallels to our leaders’ faith in computer technology as educational panacea are unmistakable.

The full report of the National Research Council has been published as *High Stakes: Testing for Tracking, Promotion, and Graduation* by the National Academy Press. A short version, and information about ordering the book, can be found at <http://www.nap.edu/readingroom/books/highstake>

Edward Miller is an educational researcher, writer and policy analyst. He is the former editor of *The Harvard Education Letter*.

Ed. Note: The above contribution is taken from NETFUTURE, an online newsletter regarding issues of technology and human responsibility. We include it here as it provides a context for the questions of assessment, learning expectations and evaluation which are currently being discussed within the Waldorf education movement. We hope to include further contributions on this subject in the January ‘00 issue of the *Bulletin*. NETFUTURE is edited by Stephen Talbott, and can be reached at <http://www.oreily.com/~steve/netfuture>