Reprint Series
1 December 1995, Volume 270, pp. 1446–1448

SCIENCE

Myths About Test Score Comparisons

Iris C. Rotberg

Myths About Test Score Comparisons

Iris C. Rotberg

In recent years, our expectations about what we can learn from testing students have become increasingly unrealistic. We use tests for inappropriate purposes and draw inaccurate conclusions from the results. To fix the perceived problem—low test scores—we administer more tests. In the process, we ignore real problems.

Testing has become an integral part of the public policy dialogue about major national issues. Scores on standardized tests are blamed for perceived failures in our economy and in international competition. They drive the debate on school reform. When educators express concern about the focus on standardized tests, we create new and, inevitably, more time-consuming tests that do not address the basic problem: Test score comparisons are highly misleading indicators of the quality of education and are irrelevant to decisions about the wisdom of any particular school reform.

I will address here a set of myths that surround standardized testing. Let me acknowledge at the outset, however, that tests can be valuable for some purposes. They have been used effectively to measure student progress, predict future performance, diagnose learning problems, encourage changes in curriculum and teaching methods, and describe national trends. However, the current use of tests has gone well beyond the reality of what they can accomplish.

Myth 1: Test score comparisons between nations, states, or schools provide valid measures of the quality of education. The international science and mathematics comparisons demonstrate the fallacy of equating test scores with school quality. These comparisons are methodologically flawed and have little to do with the quality of education. The basic problem is student selectivity: The fewer the students who take the test, the higher the average score. That score is not a valid measure of the overall quality of the education system. It simply reflects the fact that the students represented in the test comparisons have been much more highly selected in some countries than in others.

In addition, the test results reflect differences among nations in the proportion of low-income children in the test-taking population. The United States, for example, has a large proportion of low-income students as compared with many other industrialized countries. There is a strong association between poverty and low test scores. We tend to hold the education system responsible for test results reflecting broad societal problems.

Test score rankings also reflect differences in curriculum emphases among nations; for example, the proportion of 12th-grade students who study calculus, the degree of subject-matter specialization after age 16, and the amount of time devoted to cram courses in addition to regular schooling. The decision about whether or not to adopt a particular educational practice should be based on a careful consideration of the merits of the proposed change, not on rankings on standardized tests that compare quite different systems (*1*).

Sampling problems found in international studies also apply to state rankings on the Scholastic Aptitude Test (SAT). The states with the highest proportions of students taking the SAT tend to have the lowest average SAT scores (*2*).

Comparisons of schools within a school system are similarly biased by sampling problems. The fewer and more highly selected the students who take the test, the higher the average score. That score has little to do with the quality of the school.

Schools can raise their scores by excluding low-performing students. After an elementary school was put on probation for low test scores, the third graders made major gains within a single school year because the "officials simply stopped testing most of the third graders. . . . [Four years later], only 28 percent of the class took the standardized test. . ." (*3*).

Schools also inflate their scores by encouraging students to drop out of school before the examination or by retaining them in their grade. An educator put it this way: "I'm concerned because we have fewer students after grade 9 and it looks like it's to a school's advantage to get a kid to drop out rather than to keep him on the rolls and have poor test scores at grade 12" (*4*).

This technique is not limited to the United States. A World Bank study described primary schools in Kenya that increased test scores by encouraging low-achieving students to drop out before the test was administered. And as many as 20% of Chinese students may be retained in grade in upper-middle school in order to increase that school's scores—and, therefore, its reputation—on university entrance examinations (*5*).

Myth 2: The quality of our schools has declined; that is why we are no longer competitive. We incorrectly conclude from the flawed test comparisons that our schools or our parents or our students have failed. We overestimate the quality and rigor of education in previous generations. We ignore the strides that have been made in educating a large proportion of the population. In 1940, 38.1% of 25- to 29-year-olds in the United States had graduated from high school. By 1993, that percentage had risen to 88.2%. In the same time period, graduation rates from 4-year colleges rose from 5.9% to 23.7% (*6*). Moreover, our educational accomplishments equal and in many cases surpass those of previous years. A recent study by the RAND Corporation found that students' reading and mathematics performance improved for all racial and ethnic groups between 1970 and 1990 (*7*).

Clearly, the United States faces serious educational problems, but they are not the problems identified by the public rhetoric. In the 1950s, we responded to Sputnik by blaming the schools for a perceived inferiority to the Soviet Union in science and technology. Later, we predicted a shortage of scientists and engineers in the 1990s—again due to the failures of our education system. Both concerns were unjustified.

We continue, however, to hear about problems in international competitiveness. The conventional wisdom is that U.S. economic competitiveness has declined because our schools produce a poorly trained work force. Yet, the evidence shows that the problems are caused by quite different factors, such as the realities of the global economy, business practices, and government policies—for example, financial incentives that encourage offshore manufacturing; differential wage rates, profit margins, and government subsidies; licensing practices; exchange rates; and trade policy.

Myth 3: We can fix our schools by administering more tests. Or, if we hold teachers accountable for students' standardized test scores our schools will improve. The evidence shows the opposite.

Testimony before the U.S. House of Representatives put it this way: "[Test-based accountability] has been tried many times over a period of centuries in numerous countries, and its track record is unimpressive. . . . It was the linchpin of the educational reform movement of the 1980s, the failure of which provides much of the impetus for the current wave of reform. . . . Holding people accountable for performance on tests tends to narrow the curriculum. It inflates test scores, leading to phony accountability. It can have pernicious effects on instruction, such as substitution of cramming for teaching. . . . It can adversely affect students already at risk—for example, increasing the dropout rate and

The author is at 7211 Brickyard Road, Potomac, MD 20854, USA.

producing more egregious cramming for the tests in schools with large minority enrollments" (8).

Test comparisons do not provide a valid basis for an accountability system. The results do not control for changes in student population, for incentives to encourage certain students to take or not take the test, or for consistency between the test and the instructional program. We can raise test scores if we teach to the test or if we exclude low-achieving students from taking the test, but the higher scores gained under those circumstances do not reflect improved education.

The RAND study referred to above concluded: "Comparisons of simple, unadjusted test scores from one year to the next or across different schools or districts do not provide a valid indicator of the performance of the teachers, schools, or school districts unless the differences in scores are very large compared to what might be accounted for by changing demographic or family characteristics. This is rarely the case; so, any use of unadjusted test scores to judge or reward teachers or schools will inevitably misjudge which teachers and schools are performing better" (9).

A key question is whether we can alleviate the problem by using alternative measures, such as attendance rates, graduation rates, or the proportion of students going to college. Clearly, these measures provide communities with valuable information about educational accomplishments and problems. However, they do not provide an equitable basis for measuring teacher accountability. The basic problem remains: The effects of teacher quality cannot be separated from the wide range of other factors that influence school outcomes.

Myth 4: The problems in current standardized testing programs can be solved by development of new and improved tests. It is argued that innovative tests, called performance tests or portfolio assessments, will take care of flaws in current testing programs. However, little attention is paid to how long such tests take to develop, how much they cost, whether they can be administered on a large scale, the amount of instructional time they displace, and the validity of the resulting comparisons.

Studies of state testing programs show that the new tests do not reduce methodological problems, they increase them. The scoring is unreliable and measures of validity (for example, whether the tests predict students' future academic performance) are lacking (*10, 11*). Some state testing programs have tried to use complex statistical formulas to control for student background variables that might affect scores. The attempt has not worked. Indeed, it has resulted in a scoring system that is incomprehensible even to educators working within the system (*12*).

Although the new tests may draw teachers' attention toward writing and problem-solving skills and away from rote learning, this benefit could be obtained by incorporating performance tests or portfolio assessments into a school's instructional program without attempting to make comparisons that provide spurious information.

Moreover, the testing programs are extremely costly and time consuming. Researchers estimate the potential cost of national testing in five subject areas in only three grades to be more than $3 billion per year (*13*). In Kentucky's testing program, fourth-grade teachers were "overwhelmed" by the administration and grading of writing and mathematics portfolios (*14*). In Vermont, teachers spent an average of 30 hours per month, excluding training, working on mathematics portfolios—time taken from instruction children otherwise would receive (*10*).

Perhaps the best example of what happens to testing programs comes from England. In 1988, Parliament mandated national curricula and assessments. The assessments of 7-year-olds took 2 to 4 weeks out of the school year. The marking and reporting form for 14-year-olds in mathematics was 112 pages long. As a result, teachers, with strong parental support, boycotted administration of the tests and reporting of test scores. They cited a range of concerns similar to those emerging from testing programs in the United States—overwork, bureaucracy, disruption of regular schooling, flawed tests, invalid comparisons of schools, and opposition to a national curriculum (*15*). The program has been abandoned.

Myth 5: We can compensate for the inadequate resources spent on poor children by increasing testing requirements. Or, put another way, money does not matter. Research shows, however, that per pupil expenditure, teacher expertise, and class size do make a difference in student achievement (*16*). Increasing testing requirements does not buy better teachers or the attention children can receive in small schools or classes. Tests do not provide low-income inner city or rural students with science laboratories, computers, or decent facilities, amenities that affluent students take for granted.

Nor will tests reduce school finance inequities that relegate low-income children to the most poorly funded schools. For example, the 100 poorest districts in Texas spend an average of just under $3000 per student. The 100 wealthiest districts spend about $7200 per student. In Illinois, school districts spend between $2400 and $8300 per student (*17*).

We cannot improve our schools by giving more tests. The danger is that myths about testing will lead to policies that are irrelevant and counterproductive in addressing the nation's most pressing educational problems: the large proportion of children who live in poverty and the vast differences in educational resources between rich and poor schools. My greatest concern is that a focus on test scores takes attention away from our most troubled schools, the work that needs to be done to resolve the problems, and the resources needed to do it.

REFERENCES AND NOTES

1. The first set of international comparisons, conducted 25 years ago, did not take into account the percentage of the age group actually enrolled in upper secondary school. At the time the tests were administered, only about 20% of the age group in Europe attended upper secondary school—the highest achieving 20%—compared with 80% of the age group in the United States (*18*). More recent studies have tried to deal with the sampling problem by testing only those 12th-grade students who are in an academic track and taking mathematics or advanced science (*19*). These changes do not address the problem. Consider, for example, a recent assessment of mathematics students in Hungary and England. Hungary ranks near the top in the eighth-grade comparison. By the 12th grade, when Hungary retains more students in mathematics than any other country, Hungary ranks among the bottom countries. England, by contrast, scores in the bottom half in most of the eighth-grade comparisons, but ranks among the top countries by the 12th grade, when only a highly select group of students there takes the test, students who have studied science and mathematics almost exclusively since the age of 16. When a country's rank can change so dramatically between the eighth and 12th grades, it simply shows that the test comparisons are meaningless as a measure of school quality (*20*). Another example of the methodological problems comes from a comparison of 11th-grade students at several different sites: Minneapolis, MN; Fairfax County, VA; the province of Alberta in Canada; Beijing, China; Taipei, Taiwan; and Sendai, Japan (*21*). Clearly, these sites are not representative of their nations as a whole, nor were the students selected within each site representative of the age group in their communities. The comparison between China and Japan shows how biases in the sample lead to misleading findings. China ranked first even though we know that Japan educates a much higher proportion of its young people, and Japanese students often spend up to 20 hours a week in cram courses in addition to their regular schooling. The reality is that the test score rankings reflected student selectivity, not the overall performance of the education system. Like many other developing countries with scarce resources, China has an elitist education system that provides upper secondary education to only a small proportion of its young people. Although most Japanese students complete high school, a majority of Chinese students already have left school by the 11th grade. As a result, only a small proportion of the age group in China is represented in the test results. They are the highest achieving students, in the capital city, in a country with particularly wide disparities between urban and rural education. A recent study showed no significant difference between U.S. and Chinese ninth-grade scores when students were selected from both urban and rural areas. Although these samples are more representative than those in the study described above, selectivity remains a problem because a large number of Chinese students have already left school by the ninth grade and therefore are not tested (*22*).
2. *College-Bound Seniors: The Class of 1990* (College Board, New York, August 1990); I. C. Rotberg, "By

Delta Kappan **65**, 10 (1984).
3. B. Ziatos, "Scores That Don't Add Up," *New York Times*, 11 November 1994, p. 28.
4. R. F. Elmore, C. H. Abelmann, S. H. Fuhrman, paper presented at the Brookings Institution conference on Performance-Based Approaches to School Reform, Washington, DC, 6 and 7 April 1995, p. 34.
5. V. Greaney and T. Kellaghan, *Equity Issues in Public Examinations in Developing Countries*, Technical Paper No. 272 (World Bank, Washington, DC, 1995).
6. *Digest of Education Statistics: 1994*, National Center for Education Statistics 94-115 (U.S. Department of Education, Washington, DC, 1994), p. 17.
7. D. W. Grissmer, S. N. Kirby, M. Berends, S. Williamson, *Student Achievement and the Changing American Family*, MR-488-LE (RAND Institute on Education and Training, Santa Monica, CA, 1994).
8. D. M. Koretz, G. F. Madaus, E. Haertel, A. E. Beaton, *National Educational Standards and Testing: A Response to the Recommendations of the National Council on Education Standards and Testing*, CT-100 (RAND Corp., Santa Monica, CA, 1992), p. 9.
9. D. W. Grissmer, *ibid.*, p. XXXV.
10. See, for example, D. Koretz, B. Stecher, S. Klein, D. McCaffrey, *The Vermont Portfolio Assessment Program*, RP-366 (RAND Corp., Santa Monica, CA, 1995).
11. See, for example, R. K. Hambleton *et al.*, *Review of the Measurement Quality of the Kentucky Instructional Results Information System, 1991–1994* (Office of Educational Accountability, Kentucky General Assembly, Frankfort, KY, June 1995).
12. R. F. Elmore, C. H. Abelmann, S. H. Fuhrman, paper presented at the Brookings Institution conference on Performance-Based Approaches to School Reform, Washington, DC, 6 and 7 April 1995.
13. D. M. Koretz, G. F. Madaus, E. Haertel, A. E. Beaton, *National Educational Standards and Testing: A Response to the Recommendations of the National Council on Education Standards and Testing*, CT-100 (RAND Corp., Santa Monica, CA, 1992).
14. R. F. Elmore, in (*11*), p. 43.
15. "Massive Teacher Boycott Derails British National Tests," reprinted from *FairTest Examiner* (Summer 1993).
16. See, for example, R. F. Ferguson, *Harvard J. Legis.* **28**, 465 (1991); L. V. Hedges, R. D. Laine, R. Greenwald, *Educ. Res.* **23**, 3 (1994); R. J. Murnane, *Harvard J. Legis.* **28**, 457 (1991).
17. I. C. Rotberg and J. J. Harvey, *Federal Policy Options for Improving the Education of Low-Income Students, Volume I: Findings and Recommendations*, MR-209-LE (RAND Corp., Santa Monica, CA, 1993).
18. T. Husen, *Phi Delta Kappan* **64**, 7 (1983).
19. *The Underachieving Curriculum: Assessing U.S. School Mathematics from an International Perspective* (International Association for the Evaluation of Educational Achievement, University of Illinois, Champaign, IL, 1987).
20. Adapted from a more detailed analysis by I. C. Rotberg in *Phi Delta Kappan* **72**, 4 (1990).
21. *Stevenson Study of Mathematics Achievement in Alberta Schools*, Summary Report based on a presentation by Harold Stevenson, Alberta Education Department, Edmonton, Alberta, Canada, 24 October 1993.
22. J. Wang, paper presented at the 76th Annual Meeting of the American Educational Research Association, San Francisco, CA, 18 to 22 April 1995.
23. This paper is based on remarks presented at the North Central Alberta Teachers' Convention, Edmonton, Alberta, Canada, 8 to 10 February 1995, and the American Youth Policy Forum, Institute for Educational Leadership, Washington, DC, 2 June 1995.