A PREOCCUPATION WITH TEST-SCORE RANKINGS

Iris C. Rotberg
Research Professor of Education Policy
Graduate School of Education and Human Development
The George Washington University

Washington, D.C.
February 7, 2001

A PREOCCUPATION WITH TEST-SCORE RANKINGS

Iris C. Rotberg

When a nation ranks low in international science and mathematics test-score comparisons, the general consensus is that its schools have failed. Consider, for example, the following responses to the rankings of the United States in the Third International Mathematics and Science Study (TIMMS) conducted at the conclusion of secondary school. President Clinton stated that "there is something wrong with the system." U.S. Secretary of Education Riley expressed concern that Americans would not "continue to be global competitors in the new knowledge economy." Gerry Wheeler, executive director of the National Science Teachers Association, put it this way: "This study is a wake-up call for us to change the culture in the classroom." Bruce Alberts, President of the National Academy of Sciences concluded: "These results have all the elements of an education tragedy ... ."

My concern is that these conclusions are based on studies that tell us little about the quality of education in any of the participating countries and give no guidance about how to design effective education programs. In my remarks today, I will discuss the methodological problems that invalidate any attempt to link international test-score rankings to the quality of a nation's schools and then show that the same problems occur in test-based accountability programs within the United States. My conclusions do not negate the fact that some of our schools face serious educational problems. The point is that test-score rankings are an unreliable basis either for assessing educational quality or for formulating public policy.

International Comparisons

Test-score rankings provide little information about the strengths and weaknesses of schools because countries differ substantially in a wide range of variables that the studies cannot account for or control. We have little information, for example, about:

- The characteristics of participating and nonparticipating schools and students;
- The proportion of low-income students in the test-taking population;
- The extent to which the students taking the test represent a highly selected population;
- The practice with respect to the inclusion or exclusion of low-achieving students, language-minority students, students with disabilities, apprenticeship programs, and entire regions of the country;
- The mix of public and private schools, comprehensive and specialized schools, and academic and vocational schools;

- The consistency between the educational program and the test; and
- Differences in coaching practices, tracking, and school-completion rates.

These variables, which differ significantly among countries, are so confounded that we cannot know how they interact or how they affect the rankings. Test-score rankings are interpreted as an indicator of school quality when, instead, they represent the impact of a wide range of uncontrolled variables. Some countries might rank high, for example, because they have relatively small proportions of low-income students in the test-taking population. Others might rank high because their language-minority students or students in special education do not participate in the test. Countries might do well in the comparisons because they have a high drop-out rate and therefore only the higher-achieving students remain in school to take the test. Indeed, some countries might rank high because they have excellent schools—or conversely, they might rank high in spite of inadequacies in their educational systems, which are overcome by other variables.

The point is we do not know what combination of factors has contributed to each country's placement. For political and practical reasons, it is not feasible to collect the type of data that would be needed to interpret the rankings. It is one thing to design a research plan and another to implement it in the real world.

Why do these methodological problems matter? I believe they matter because policymakers make decisions based on the test-score rankings. The rankings lead to "quick fixes." We all have heard proposed "solutions" to low international test scores: Increase testing in elementary and secondary school; require more high school calculus; replace public schools with vouchers; select a high-ranking country and use its textbooks; give all students a standard curriculum; or, conversely, create more highly specialized schools. Clearly, there is room for an analysis of each policy on its merits, and we can learn from the experience in other countries, but the issues cannot be resolved by examining international test-score rankings.

Test-Based Accountability

There appears to be widespread agreement on both sides of the political aisle that student scores on standardized tests are valid measures of the quality of education in a state, a school district, or a school. Indeed, test-based accountability plans—with their associated rewards and sanctions--were the centerpiece of both presidential candidates' education proposals during the campaign. There is little disagreement about the goal. Everyone wants expert teachers and strong education programs. The issue is how to attain that goal. At first glance, the accountability plans seem quite reasonable. However, they remain controversial among researchers, educators, and parents—and for good reason.

While the test-based accountability plans are intended to improve education, I believe there is evidence that they have, instead, been counterproductive.

First, test-score rankings do not tell us which states, school districts, or schools are doing a good job. Therefore, rewards and sanctions administered on the basis of these rankings are based on flawed measures of performance. There are several reasons why standardized test-scores tell us little about the strengths and weaknesses of schools.

- There are large differences in student selectivity. That is, we do not know which students took, or did not take, the test. We do not know which students are assigned to special education programs and therefore did not take the test. We do not know whether language minority students have taken the test or, if they did, whether their scores are included. There are major differences in policies between states, between school districts, and even between schools within the same district.

- When we read the test-score results, we do not know the grade retention policies or the drop-out rates in specific school districts and schools. For example, if more students are retained in ninth grade, the average tenth grade score will go up, but more students will drop out of school. The higher scores do not mean the school has improved. They simply mean that the lowest achieving students are no longer included in the test results—and that many of these students have dropped out of school. Conversely, a similar school that tries to keep most of its students in school will have lower average test scores. That does not mean the school is inferior. It simply means it is attempting to serve a wider range of students. My example is not hypothetical, nor is it limited to the current generation of accountability plans, or to the United States. In the 1940s, Irish schools responded to accountability pressures by increasing grade retention. More recently, World Bank studies report exclusions in China and Kenya. Similar reports are emerging in the United States, for example, from Kentucky and Texas, states that place strong emphasis on test-based accountability. An assessment coordinator in Kentucky 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."

- In addition to student selectivity, test scores will be determined by many other factors—cramming for the test; test familiarity; the difficulty of the test and whether the test becomes easier or harder over the years; cheating in some cases; and, most important, poverty and the resources available to schools. Administering more tests will not overcome the fact that poverty is the major factor contributing to low educational achievement—in the United States and throughout the world. In most studies, it accounts for 75% of the

variance in student achievement scores. Consider, for example, a ranking of states by NAEP (National Assessment of Educational Progress) scores: A state's ranking on NAEP is determined to a great extent by the proportion of children living in poverty in that state. My comments do not mean that children from low-income families cannot achieve in school. Many overcome the odds and excel. Nor does it mean educators should be relieved of the responsibility to provide these children with a quality educational experience. But it does mean that if a problem is that big you need a major investment to begin to address it in any serious way. Our system of school finance does just the opposite: It compounds the educational problems associated with poverty by creating major school finance inequities, which affect everything from teacher quality to class size to course offerings. The federal contribution currently is much too small, and much too untargeted, to compensate for these inequities.

- And, even if none of these problems in interpreting test scores existed, we might question whether the federal government can—or should be—in the business of monitoring test scores in each of the 85,000 schools in the United States.

In addition to considering whether standardized tests are valid measures of the performance of principals and teachers, we can ask whether the process itself might improve the education program—that is, will it raise academic standards?

On the positive side, I hear reports of increased program focus or more emphasis on writing, if writing is part of the test. Some feel that the test reports in themselves might attract more resources to high poverty schools. But there also are major concerns about negative academic effects resulting from an emphasis on test scores.

- Many schools spend weeks, even months, on test-preparation activities. That is, the test becomes the curriculum, which replaces the school's ongoing academic program. The focus on testing, in turn, narrows the curriculum and encourages rote learning. When we read that states have raised academic standards, all we know is that they have initiated a high stakes testing program. We know nothing about whether the quality of the education program has improved. For example, if 25% of students drop out of school because they failed the test, we have not improved our schools—they simply are not serving the lower-performing students.

- In addition, a preoccupation with high stakes testing may have a negative impact on the teaching environment. The risk is that the most qualified teachers and principals will be discouraged from entering and remaining in the profession, particularly in low-income schools. There are reports of teachers leaving the field, or requesting transfers to a grade that is not tested, because they feel that the tests have adverse effects on instructional methods

and working conditions. It also is becoming increasingly difficult to attract and retain principals. *The New York Times*, reporting on shortages of principals, described it this way:

"As the academic year begins for the nation's 53 million students, a growing number of schools are rudderless, struggling to replace a graying corps of principals at a time when the pressure to raise test scores and other new demands have made an already difficult job an increasingly thankless one. ... In Kentucky and Texas, where the pace at which principals are fleeing is as accelerated as it is in Vermont, job openings in some districts that drew more than a dozen applicants as recently as five years ago are now attracting as few as three, according to principals' associations there."

If policies intended to strengthen academic standards exacerbate current shortages, they will have precisely the opposite effect from that intended.

- High stakes testing also weakens the quality of education if it encourages policies that may not be in the best interest of the child--for example, policies, described earlier, that increase drop-out rates, or decrease graduation rates.

Finally—and most troublesome—is the fact that the focus on test-based accountability has diverted attention from underlying causes of low academic achievement. We cannot improve education for "all" children without addressing problems of poverty and the serious inequalities in resources available to schools serving affluent and low-income populations. Nor can a test substitute for a comprehensive and sustained academic program or a working environment that encourages the most qualified teachers and principals to remain in the profession.