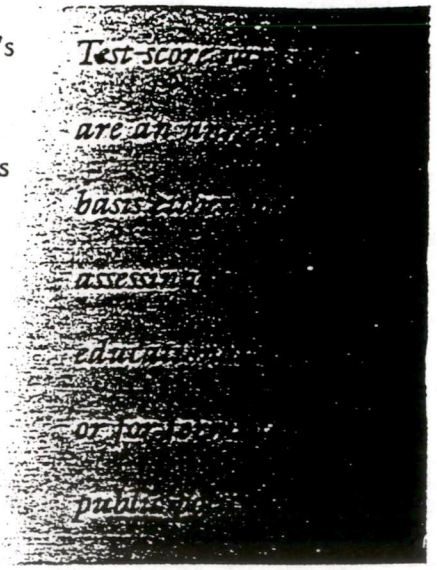
Do Flawed International Test-Score Comparisons Matter?

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 relatively low rankings of the United States in the Third International Mathematics and Science Study (TIMSS) conducted at the conclusion of secondary school. President Clinton stated that "there is something wrong with the system." U.S. Secretary of Education Richard W. 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."

My concern is that these conclusions are based on a misleading and seriously flawed study that tells us little about the quality of education in any of the participating countries and gives no guidance about how



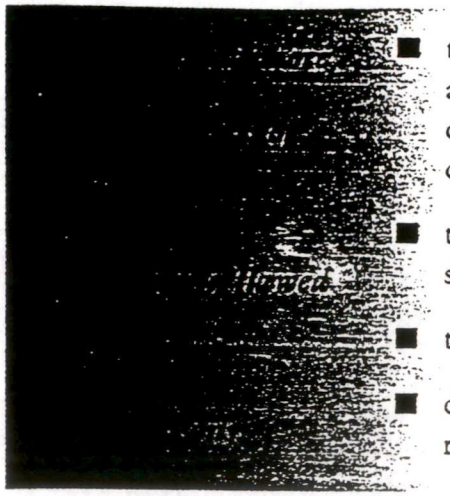
to design effective educational programs. Here, I discuss the study's methodological problems and, in turn, the costs of continuing to report invalid findings. My conclusions do not negate the fact that some of our schools face serious educational problems. The point is that test-score rankings are an unreliable basis either for assessing educational quality or for formulating public policy.

Methodological Problems

Test-score rankings provide little information about educational quality because countries differ substantially in a range of variables that the international comparisons cannot control — for example, the proportion of the age cohort participating in the test, the proportion of low-income students in the test-taking population, and the age of the participating students, which ranged from 17 to 21 in TIMSS. The failure of most countries to meet TIMSS' sampling standards exemplifies the problem. Although only five of the 21 countries participating in the mathematics and science general achievement tests and only six of the 16 countries participating in the advanced mathematics and physics tests met the standards for selecting schools and students, their results were incorporated into the rankings. Low participation and high exclusion rates tend to increase a country's rank because lower achieving schools and students are more likely to be excluded from the testing program.

To show a relationship between the test scores and educational quality in each participating country, we would need to identify and then measure the potential confounding variables. The international studies do not have the data to conduct the type of multivariate analyses required to make a systematic assessment of the impact of uncontrolled variables, and it is not likely that these data will be available in future studies. The problems inherent in making valid comparisons even within a single school district or state in the United States are well documented. These comparisons become even more unrealistic when we attempt to attribute test scores to school quality across a set of diverse nations from which we have little information about a broad range of variables including, for example:

- the characteristics of participating and nonparticipating schools and students;
- 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;
- differences in coaching practices, tracking, and school-completion rates; and
- variations in family socioeconomic status and how it is measured in different countries.

These variables, which differ significantly among countries, are so confounded that we cannot know how they interact or how they affect the rankings. There are enormous practical and political obstacles in acquiring quality data on the confounding variables and in controlling for these variables even when data are available. As a result, we have had more than 30 years of experience in conducting flawed test-score comparisons — sufficient evidence, I believe, to demonstrate that the prognosis for future studies is not good. It is one thing to design a sampling plan and another to implement it in the real world.

Negative Consequences of Conducting Flawed Comparisons

A concern about the validity of international test-score comparisons goes beyond an academic discussion of research design and multivariate analysis. I believe these comparisons matter for several reasons:

The test-score comparisons mislead by producing unsubstantiated and highly publicized conclusions about the relative quality of education in participating countries.

The test-score rankings are interpreted as an indicator of school quality when, instead, they are likely to represent the impact of a range of uncontrolled variables. The rankings provide no information about which factors contribute to a country's placement. Countries might rank high in mathematics because low-

achieving students were not in the test-taking population, because the students who took the test attended highly specialized science and technical schools, or because students from low-income regions were excluded from the study. 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. Consider, for example, the following comparisons:

- In the Czech Republic, the combined participation rate of schools and students was 92 percent, the average age of the participating students was 17.8, and a range of programs and grades was represented. Denmark, in contrast, had a participation rate of 49 percent and an average age of 19.1, and excluded all students from testing who had only nine years of formal schooling.

- Italy had a graduation rate of 49 percent (therefore, only the "top" half of students was eligible to take the test), excluded an additional 30 percent of these eligible students from the population to be sampled, and had a relatively high poverty rate. Sweden, however, had a graduation rate of 88 percent, included all eligible students in the sampling frame, and had fewer students living in poverty.

- Latvia tested students only in physics, had a 50 percent exclusion rate, and represented only 3 percent of the age cohort in the physics assessment. Austria, in contrast, tested students in all components of the study, had an exclusion rate of 18 percent, and represented 33 percent of the age cohort in the physics assessment.

The comparisons, therefore, give us virtually no information about the relative strengths and weaknesses of education in these countries.

The test-score comparisons lead to "quick fixes" because they are not designed to provide information that is useful in formulating education policy.

We have all heard proposed "solutions" to low test scores: increase testing in elementary and secondary school; require more high school calculus; replace public schools with vouchers; 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, but the issues cannot be resolved by examining international test-score comparisons.

Apart from their methodological flaws, these comparisons ignore the broader context by focusing on a single measure — scores on standardized tests. Alternative criteria in evaluating the quality of the educational experience would provide a much better frame of reference for assessing our strengths and weaknesses and for formulating public policy.

The test-score comparisons trivialize international research and detract from potentially valuable studies.

An approach to international research that focuses on the benefits and shortcomings of alternative educational practices rather than on test-score rankings would provide information directly relevant to policy deliberations. The research might address key issues in science and mathematics education, as well as more generic issues that countries must consider in setting education policy and assessing tradeoffs — for example: resource allocation; the status, training, and working environment of teachers; assessment and accountability procedures; curriculum policies; and access to higher education. Although research exists on many of these topics, including some material that is part of the current TIMSS report, it typically is not designed specifically to inform public policy and certainly does not receive the same attention as do test-score rankings.

Let us celebrate the turn of the century by conducting international studies that can serve to strengthen our educational system rather than by continuing unproductive and flawed research.

Iris C. Rotberg *is research professor of education policy in the Department of Educational Leadership, Graduate School of Education and Human Development, The George Washington University, Washington, D.C. This paper is adapted in part from an article that appeared in the May 15, 1998, issue of* Science.