POLICY: SCIENCE EDUCATION

Interpretation of International Test Score Comparisons

Iris C. Rotberg

The most recent findings of the Third International Mathematics and Science Study (TIMSS) (*1*) prompted widespread concern because the United States ranked relatively low in test score comparisons made at the end of secondary school. These reactions are based on a misleading and seriously flawed study. The methodological difficulties found in earlier studies have not been alleviated in this one. TIMSS, like its predecessors, tells us little about the quality of education in any of the participating countries and provides no guidance about how to design effective education programs. Here, I discuss the methodological problems in the comparisons and suggest an approach to international studies that would increase their relevance.

Test score rankings provide little information about educational quality because countries differ substantially in such factors as student selectivity, curriculum emphases, and the proportion of low-income students in the test-taking population (*2*, *3*). Like its predecessors over the past 30 years, the current study has not controlled for these factors. Although the executive summary of the TIMSS report assures the reader that "the students who participated in TIMSS were scientifically selected to accurately represent students in their respective nations" [(*1*), p. 13], the actual data presented in the body of the report are less reassuring.

Sampling

TIMSS tested students at the end of secondary school in mathematics general achievement, science general achievement, advanced mathematics, and physics (see tables at www.sciencemag.org/feature/data/981368.shl).

Most of the participating countries failed to meet the TIMSS sampling standards for selecting schools and students—many by a substantial margin. Only 5 of the 21 countries participating in the mathematics and science general achievement tests and only 6 of the 16 countries participating in the advanced mathematics and physics tests met the "international sampling and other guidelines" established by TIMSS.

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. Indeed, the very reason that TIMSS provided the guidelines, which were "more honor'd in the breach than the observance," was to prevent that occurrence from influencing the reported rankings.

"Eligible" Populations

The higher the proportion of the age group who take the test, the lower will be the average score. That is why, for example, the U.S. states with the highest proportions of students taking the SAT tend to have the lowest average SAT scores Those scores reflect student selectivity, not the quality of education (*4*).

Similarly, countries with relatively few students taking the test also can be expected to score higher. If a country has a low graduation rate, its average test scores will tend to be inflated because lower achieving children, who have already left school, are not tested. The TIMSS report recognized the problem but concluded that it did not apply to the study [(*1*), p. 13]. However, the problem was not solved. Although U.S. enrollment and graduation rates are similar to the average rates of participating countries that reported statistics, there were significant gaps among individual countries. For example, the percentage of 25- to 34-year-olds who had completed secondary education varied from 49% in Italy to 91% in the Czech Republic. Seven participating countries did not report graduation rates.

Countries also varied substantially in the proportion of the age group taking the advanced mathematics and physics tests. Only students who had taken advanced courses in these areas were eligible to take the tests. Therefore, countries with a high proportion of students taking advanced courses are at a disadvantage. For example, the percentage of the age cohort represented in the advanced mathematics test varied from 2% in the Russian Federation, 3% in Lithuania, and 9% in Cyprus to 26% in Germany, 33% in Austria, and 75% in Slovenia.

Age and Grade

TIMSS tested students in their final year of secondary education. In some countries, the final year was after 10 years of schooling; in others, it was after 14 years. As a result, the age of students taking the general knowledge assessments ranged from 17 to 21.

The average age of the students taking the test in each country clearly influences the country's rank, as well as its relative performance between eighth grade and the final year of secondary school. In the general mathematics assessment, five countries ranked higher in the final year than in eighth grade, six (including the United States) ranked lower, and nine maintained their position. Some observers have interpreted this decline in the U.S. position as an ominous indicator of the failure of our education system. However, the TIMSS analysis points out that the countries that declined had the smallest average age gap between the two grades (3.5 years), whereas those that gained had the largest age gap (5.4 years). The finding simply shows that those who scored higher in the final year of secondary school were older, more advanced students. It tells us nothing about whether there was a deterioration in the quality of schools between eighth grade and the final year of schooling.

Type of School and Poverty

The TIMSS report states that "the strict quality controls ensured that the sample of students taking the general knowledge assessments was representative of all students at the end of secondary school, not just those in academically-oriented programs" [(*1*), p. 13]. However, some countries tested a range of diverse schools, whereas others excluded vocational schools, apprenticeship programs, or private schools.

Differences between types of school were particularly pronounced in the advanced mathematics and physics assessments. Students in some countries attended highly specialized schools or programs, which attract the highest achieving students and focus primarily on science and mathematics. In Cyprus, students taking the advanced mathematics test were in their final year of the mathematics and science program; in France, the final year of the scientific track; in Lithuania, the final year of the mathematics and science gymnasia; in Sweden, the final year of the natural science or technology lines; and in Switzerland, the final year of the scientific track of gymnasium. In contrast, students in several countries, including the United States, attended comprehensive secondary schools. The major differences in student selectivity and school specialization across countries make it virtually impossible to interpret the rankings. Nor do the test score comparisons provide

The author is in the Department of Educational Leadership, Graduate School of Education and Human Development, The George Washington University, Washington, DC 20052, USA. E-mail: irotberg@gwis2.circ.gwu.edu

information about the wisdom of specialized schools—the advantages, or disadvantages, of encouraging young students to make specific career choices, or the impact of tracking on the general student population.

A large body of research has demonstrated that there is an association between low student achievement and poverty (5). A country's rank will be influenced by its proportion of low-income children in the test-taking population. The countries participating in TIMSS differ significantly in poverty levels (6), and there can be little doubt that poverty and its associated societal problems—crime, violence, poor health and nutrition—played a significant role in the TIMSS findings. The study was not designed to make that analysis possible.

Cumulative Effects

Each of the methodological problems influences the international test score rankings. TIMSS did not have the data to conduct the type of multivariate analyses required to make a systematic assessment of their impact of uncontrolled variables, and it is not likely that these data will be available in future studies (*7*, 8). The variables, which occur to different degrees in different countries, are so confounded that we cannot know how they interact or how they affect the rankings.

We do know, however, that countries had such different patterns of participation and exclusion rates, school and student characteristics, and societal contexts that test score rankings are meaningless as an indicator of the quality of education. For example, in the Czech Republic, the participation rate was 92%, the average age of the participating students was 17.8, and a wide range of programs and grades was represented. In contrast, Denmark had a participation rate of 49%, an average age of 19.1, and it excluded all students from testing who had only 9 years of formal schooling.

Italy had an exclusion rate of 30%, a graduation rate of 49%, and a relatively high poverty rate. Sweden had an exclusion rate of 0%, a graduation rate of 88%, and less poverty. Latvia tested students only in physics, had a 50% exclusion rate, and represented only 3% of the age cohort in the physics assessment. Austria tested students in all components of the study, had an exclusion rate of 18%, and represented 33% of the age cohort in the physics assessment.

Moreover, TIMSS does not provide the information needed to identify the research design artifacts that might have influenced a given country's relative performance across the four tests. We do know, however, that the rankings are unstable: France, for example, moved from 7th place in the mathematics general knowledge test, to 13th place in the science general knowledge test, to 1st place in the advanced mathematics test, and back to 13th place in the physics test. The Russian Federation ranked 15th and 16th in the science and mathematics general knowledge tests but moved to second and third place in the advanced mathematics and physics tests.

In short, the methodological problems of the most recent international comparisons are as great as those in previous studies. The studies are irrelevant to deliberations about educational reform or as predictors of a nation's scientific and technological strength.

Policy Implications

Thirty years of experience with international test score comparisons have shown that their flaws consistently lead to misleading findings that have little policy relevance. There are clearly alternative criteria to test score comparisons in evaluating the quality of the educational experience in a given country (3). For example:

- Productivity in science and engineering, as measured by breakthroughs in basic research, technological advances, and product development.
- Research opportunities in institutions of higher education.
- The availability of qualified scientists and engineers to meet workforce requirements.
- Retention and graduation rates in science and mathematics education.
- Participation of women and minorities in science and engineering.
- Access to higher education in science and engineering for low-income students, students from racial and ethnic minority groups, and students with disabilities.
- Equality of opportunity to participate in science and mathematics programs in elementary and secondary school, as measured by such indicators as the distribution of resources, school environment, and programs for students with special needs.
- The availability of science and mathematics education for students who do not attend college.
- An adequate supply of qualified science and mathematics teachers in elementary and secondary school.

International studies could be productively designed to identify how various countries address these issues and to evaluate the effectiveness of alternative policies. A large body of research already exists on many of the topics, including some material that is part of the current TIMSS report. The point would be to build on this work to provide systematic information about effective practices. Because the strength of a country's education system depends on broad economic and social conditions, as well as on schooling practices, the studies also might consider such variables as poverty rates and associated societal problems, income disparities, and fiscal policy.

An approach to international research that focused on the benefits and costs of alternative educational practices rather than on test score rankings could provide information directly relevant to policy deliberations. It might not make headlines, but it would provide a much stronger basis for improving education.

Additional data can be found at *Science* Online at www.sciencemag.org/feature/data/981368.shl

References and Notes

1. U.S. Department of Education, National Center for Education Statistics (NCES), *Pursuing Excellence: A Study of U.S. Twelfth-Grade Mathematics and Science Achievement in International Context*, NCES 98-049 (U.S. Government Printing Office, Washington, DC, 1998).
2. See, for example, T. Husén, *Phi Delta Kappan* **64**, 455 (March 1983); H. J. Kiesling, *Econ. Educat. Rev.* **13**, 179 (1994).
3. I. C. Rotberg, *Phi Delta Kappan* **72**, 296 (December 1990).
4. The College Board, *News from The College Board* (New York, 26 August 1997).
5. NCES, *Education in States and Nations: Indicators Comparing U.S. States with the OECD Countries in 1988*, NCES 93-237 (NCES, Washington, DC, 1993); The Annie E. Casey Foundation, *Kids Count Data Book: State Profiles of Child Well-Being* (The Annie E. Casey Foundation, Baltimore, MD, 1995).
6. See, for example, M. L. Blackburn, *Comparing Poverty: The United States and Other Industrialized Nations* (American Enterprise Institute for Public Policy Research, Washington, DC, 1997).
7. Although multivariate analysis might have improved the study, it could not overcome the basic problem. In order to carry out a multivariate analysis—and show a relationship between the test scores and educational quality in each of the participating countries—we would need to identify, and then measure, the potential confounding variables. The problems inherent in making comparisons even within a single school district or state in the United States are well documented (*8*). These comparisons become even more unrealistic when we attempt to link test scores with school quality across a set of diverse nations from which we have little information about a broad range of variables: the characteristics of participating and nonparticipating schools and students; varying practices with respect to the inclusion or exclusion of low-achieving students, language-minority students, students with disabilities, types of programs, or entire regions of the country; the mix of public and private schools; the consistency between the educational program and the test; differences in coaching practices, tracking, and school completion rates; variations in family socioeconomic status and how it is measured in different countries; and many more. There are enormous practical and political obstacles to acquiring quality data on the many variables, and for this reason international test score comparisons to date have not conducted multivariate analyses.
8. See, for example, R. F. Elmore, C. H. Abelmann, and S. H. Fuhrman, in *Holding Schools Accountable: Performance-Based Reform in Education*, H. F. Ladd, Ed. (Brookings Institution Press, Washington, DC, 1996), pp. 65-98; 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); I. C. Rotberg, *Science* **270**, 1446 (1995).