Teachers College Record
The Voice of Scholarship in Education

Tenuous Findings, Tenuous Policies

by Iris C. Rotberg – May 19, 2014

"Not everything that counts can be counted, and not everything that can be counted counts." -Albert Einstein

Methodological problems have plagued international test-score comparisons from the time they began 50 years ago. Since then the number and type of countries and other jurisdictions participating in the comparisons have increased, as have the methodological problems. At the same time, the results of the international comparisons have had an increasing impact on education policies throughout the world, despite the fact that the policy implications drawn from the comparisons are based on seriously flawed data. The commentary describes the intractable problems inherent in making valid comparisons of student achievement across countries and recommends an approach to reformulating the research.

THE PROBLEM

The methodological critiques of international test-score comparisons began shortly after the comparisons were first administered 50 years ago, and they have continued. Methodological critiques of research are not unusual, but this situation is quite extraordinary for several reasons. First, the critiques of the international test-score comparisons are extensive and address virtually every aspect of the studies—sampling, measurement, and interpretation. Second, the studies continue to be administered, with few of the critiques addressed, but with continued participation of a large number of countries and other jurisdictions. These massive data collection efforts have been conducted 13 times in the past 18 years. The results of the most recent study, the 2012 Program for International Student Assessment (PISA), were released in December 2013, only a year after the release of the other two major comparisons, the Trends in International Mathematics and Science Study (TIMSS) and the Progress in International Reading Literacy Study (PIRLS) (National Center for Education Statistics, n.d.-a, n.d.-b, n.d.-c). Third, despite the critiques, the studies have had a large impact on political rhetoric, public opinion, and public policies in countries throughout the world. This commentary focuses on PISA, the most recent international comparison released. Although the three international comparisons differ in some respects, the basic methodological problems described here are inherent in international test-score comparisons more generally.

The recent round of critiques of PISA has focused on the serious sampling problems in Shanghai. The test-score results for Shanghai, one of the richest cities in China, are included in the rankings along with results for entire countries. Moreover, even in Shanghai, the sample of children tested is not representative of the demographics of the city because most of the children of the millions of migrant workers are no longer in school at age 15 when the test is administered or have been sent back to their home provinces because of the requirements of the *hukou*, the Chinese registration system. In either case, the children are not tested (Loveless, 2013). The test results, therefore, are far from representative of the population of Shanghai, much less of China.

Shanghai is a particularly egregious example, but what is overlooked in the discussion is that Shanghai is just the tip of the iceberg. The sampling problem is endemic. First, many of the countries participating in PISA also have large numbers of children who do not attend school at age 15. In Vietnam, which is highlighted in the recent comparisons, more than a third of the age group and about two-thirds of the poorest 20% are no longer in school when the test is administered (Bodewig, 2013). Many of the other developing countries also have large numbers of children who do not attend school (Filmore, 2010). We know that lower-achieving children will inevitably be underrepresented in countries with significant proportions of children out of school and that this underrepresentation seriously biases the comparisons. It is meaningless to compare these countries either with each other or with countries that have a high proportion of children in school. Elegant sampling designs cannot compensate for the reality of children who are not represented in the test results,

Second, sampling remains a serious problem even in industrialized countries, where school enrollment rates are higher, but still vary across countries (Organization for Economic Cooperation and Development [OECD], 2013a). Moreover, information is not available to control for the differential representation of children from particular population groups—for example, children in special education, vocational education, apprenticeship programs, or remote rural areas, as well as children from racial and ethnic minority groups, children of immigrants, language-minority children, and low-income children. These children, on average, are less likely to remain in school than the general student population. They drop out of school at different rates in different countries (Eurostat, 2013; UNICEF, 2011) and, in some cases, are in special schools or classrooms, or in workplaces (World Health Organization & The World Bank, 2011; Crul, 2007), which are less likely to be tested. We do not know the extent to which these children are, or are not, represented in the test results in each of the participating countries and, therefore, whether the test results are comparable across countries.

This information is not available to assess the effects for even a single population group (for example, children in special education), and we certainly do not have the information needed to estimate the cumulative effects for the underrepresented populations more generally. Without knowing the extent or nature of the bias in each country, we cannot assess either the validity or the significance of the rankings.

TENUOUS CONCLUSIONS

The international test-score rankings are almost universally interpreted by countries as an indication of the quality of their schools, despite the extensive methodological problems that make it virtually impossible to draw causal relationships between test scores and school quality. We are taking tenuous results and applying them in a questionable way. Even if the rankings were sound, a causal leap from test-score rankings to school quality would be unwarranted given the wide range of other factors that influence the rankings, such as the differences among countries in poverty rates, income distribution, immigration rates, social support services, and the extent to which children participate in academic programs and cram courses outside of school. And beyond all of these variables, there remains the basic question of whether a test score is a fair representation of the complexity and quality of a country's entire education system. It has proven to be virtually impossible to unravel the cumulative effects of all the uncontrolled variables and then make valid interpretations of the implications of the test-score rankings.

UNPRODUCTIVE POLICIES

These concerns could be dismissed as trivial if they did not play such a major role in public policy and political rhetoric and if the policies they led to had turned out to be productive. But that is not the case.

The international test-score comparisons have had a major impact on education policies in countries throughout the world. In the United States, for example, they have contributed to the perception that U.S. public schools are failing and led to a preoccupation with test-based accountability—a policy that has had questionable results. It is ironic that this policy, initiated in part in response to U.S. ranking on international tests, is inconsistent with the policies of many other countries (including those that are admired because of their ranking), which generally do not hold teachers accountable for their students' test scores and sometimes actively discourage the practice (Bonnet, 2010; Watanabe, 2010).

Even beyond increased testing, policy analysts have drawn conclusions from the international test-score rankings that go well beyond the research evidence, and often contradict it. For example, the ranking of the United States on science and mathematics tests is taken as an indicator of future shortages of scientists, mathematicians, and engineers and, further, of an inability to compete in the global economy. Neither conclusion bears any relationship to the findings of the comparisons, even if these findings were sound. Both conclusions are also inconsistent with other research findings, including labor market projections (Lowell & Salzman, 2007) and rankings on global competitiveness (IMD World Competitiveness Center, 2013). However, both conclusions continue to have a major impact on policy deliberations.

The international test-score comparisons are a major industry. The number of participating countries and jurisdictions has increased since the tests were first administered. So has the number of researchers and educators involved in designing and administering the comparisons, writing volumes of analysis, critiquing the comparisons, and critiquing each other's critiques. The angst among educators has grown, as has the political rhetoric surrounding the comparisons. The researchers designing the test publish some caveats (and omit others), but these caveats disappear into the "cloud" or similar obscure sites. What remains is the conventional wisdom that the schools in low-scoring countries have failed, while those in high-scoring countries are just fine. None of this will change without a basic reformulation of the research approach used by the international studies.

CAN THE PROBLEM BE FIXED?

Sampling problems are endemic in any test-score comparisons, even those involving neighboring domestic school districts or states. These problems are magnified in international comparisons, where countries vary enormously in the proportion of the age group represented in the comparisons and in the representation of particular population groups. Yet, the studies treat the comparisons as if the student samples were equally representative of the population in each of the participating countries. They are not, and no sampling design can solve the problem. The results of the comparisons, therefore, are inevitably flawed. Comparing student achievement across countries by measuring only the performance of students in school is like comparing health across countries by assessing only those who go to doctors and ignoring populations who do not have access to medical care.

Even the most skilled sampling expert cannot design a study that represents populations that are not tested. Yes, the sponsors of the comparisons could be more transparent about the flawed results. They could give more information about the characteristics and percentages of excluded students. They could eliminate all countries that had large proportions of children out of school, a choice that would exclude virtually every developing country and limit the comparisons to a small group of privileged countries. That choice would be unfortunate, and it would also not solve the basic problem of comparability, which applies even to industrialized countries because of continuing differences among countries in overall school participation rates and in participation rates for particular population groups.

Rather than marginal changes, I suggest a basic reformulation of the research approach used by the international studies so the studies are no longer focused on rankings of countries by test scores, but concentrate instead on specific problem areas in individual countries or clusters of countries. This approach would have a number of advantages. Eliminating test-score comparisons would facilitate the design of methodologically sound studies that are not constrained by sampling flaws. Focusing resources on specific problem areas in individual countries, rather than on the massive data collection efforts required for test-score comparisons, would provide more opportunity to analyze each country's education system in the broader social and economic context. The studies could concentrate on the issues that are most relevant to each country and tailor the research to meet the needs of those countries. The results, in turn, would have direct implications for public policy.

A wide range of topics could be investigated including, for example, research on factors within countries that mitigate or exacerbate achievement gaps and public policies that make a difference. PISA has considered some of these questions, but the results typically build on the test-score comparisons and, therefore, have the same flaws. For example, PISA uses as a measure of equity the proportion of variance in student achievement that is accounted for by the socioeconomic status of families and schools (OECD, 2013b). When countries are ranked by this measure, here too the rankings cannot be interpreted because they provide no information about the underlying factors contributing to each country's equity rank. If a country ranks high on equity (defined as a low achievement gap based on socioeconomic status), the rank could reflect the country's low poverty rate, its equitable financing of schools, its high intergenerational mobility, its low immigration rate, or the fact that low-income children are less likely to attend school. Essentially, we do not know whether a country that is ranked high on equity is in fact equitable or whether the rank is an artifact of policies that discourage immigration or exclude large proportions of low-income children from school, which might give the illusion of high equity when just the opposite is true since the more affluent children are overrepresented in the test results.

PISA's own findings support a transition to studies of individual countries. They show that the proportion of variance in student achievement accounted for by socioeconomic status and other differences *within* member countries in the Organization for Economic Cooperation and Development (OECD) is nine times greater than the proportion accounted for by differences *among* OECD countries (OECD, 2010)—a finding that has been obscured by the emphasis on test-score rankings and largely ignored in the public dialogue. It is consistent with a research approach that focuses on problem areas within countries rather than on test-score competitions among countries. It also offers an opportunity to take Einstein's advice and focus on issues that count, and count only what can be counted. After 50 years of test-score rankings, it's worth a try.

References

Bodewig, C. (2013, December 11). What explains Vietnam's stunning performance in PISA 2012? In *World Bank blog.* Retrieved from http://blogs.worldbank.org/eastasiapacific/what-explains-vietnam-s-stunning-performance-pisa-2012

Bonnet, G. (2010). France: Diverse populations, centralized administration. In I. C. Rotberg (Ed.), *Balancing change and tradition in global education reform* (2nd ed.). Lanham, MD: Rowman & Littlefield Education.

Crul, M. (2007). *Pathways to success for the second generation in Europe.* Washington, DC: Migration Policy Institute. Retrieved from http://www.migrationpolicy.org/article/pathways-success-second -generation-europe

Eurostat. (2013). Education. In *Smarter, greener, more inclusive? Indicators to support the Europe 2020 strategy.* Luxembourg: Publications Office of the European Union. Retrieved from http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-02-13-238/EN/KS-02-13-238-EN.PDF

Filmore, D. (2010). *Education and attainment around the world: An international database.* Washington, DC: The World Bank. Retrieved from http://econ.worldbank.org/projects/edattain

IMD World Competitiveness Center. (2013). *IMD world competitiveness rankings 2013.* Lausanne, Switzerland: Author.

Loveless, T. (2013, December 11). Attention OECD-PISA: Your silence on China is wrong. In *Brookings blog.* Retrieved from http://www.brookings.edu/blogs/brown-center-chalkboard/posts/2013/12/11-shanghai-pisa-scores-wrong-loveless

Lowell, B. L., & Salzman, H. (2007). *Into the eye of the storm: Assessing the evidence on science and engineering education, quality, and workforce demand.* Washington, DC: Urban Institute.

National Center for Education Statistics. (n.d.-a). *Trends in International Mathematics and Science Study (TIMSS).* Retrieved from http://nces.ed.gov/timss/

National Center for Education Statistics. (n.d.-b). *Progress in International Reading Literacy Study (PIRLS).* Retrieved from https://nces.ed.gov/surveys/pirls/

National Center for Education Statistics. (n.d.-c). *Program for International Student Assessment (PISA).* Retrieved from http://nces.ed.gov/surveys/pisa/

Organization for Economic Cooperation and Development. (2010). *PISA 2009 results. Overcoming social background: Equity in learning opportunities and outcomes* (Vol. II). Paris, France: OECD Publishing.

Organization for Economic Cooperation and Development. (2013a). *Education at a glance 2013: OECD indicators.* Paris, France: OECD Publishing. http://dx.doi.org/10.1787/eag-2013-en.

Organization for Economic Cooperation and Development. (2013b). *PISA 2012 results. Excellence through equity: Giving every student the chance to succeed* (Vol. II). Paris, France: OECD Publishing.

UNICEF. (2011). *The right of Roma children to education: Position paper.* Geneva, Switzerland: UNICEF Regional Office for Central and Eastern Europe and the Commonwealth of Independent States (CEECIS).

Watanabe, R. (2010). Japan: Encouraging individualism, maintaining community values. In I. C. Rotberg (Ed.), *Balancing change and tradition in global education reform* (2nd ed.). Lanham, MD: Rowman & Littlefield Education.

World Health Organization & The World Bank. (2011). *World report on disability.* Geneva, Switzerland: WHO Press.