INFORMATION CAPSULE
Research Services

Vol. 0802
September 2008

Christie Blazer, Supervisor

INTERNATIONAL ASSESSMENTS: WHAT DO THEY REALLY TELL US?

At A Glance

This Information Capsule discusses the limitations associated with international assessments and advises educators to consider these limitations before drawing conclusions regarding the United States' standing in the global education community. Some researchers, in fact, have suggested that it is more effective to transfer best practices across cities and states than to adopt poorly understood practices found in an assortment of small countries around the world.

Few education stories get as much attention in the media as the performance of American students on international tests. But are these comparisons valid? What do they really tell us about the status of education in the United States? This Information Capsule discusses the limitations associated with international assessments and advises educators to consider these limitations before drawing conclusions regarding the United States' standing in the global education community.

International assessments have long been seen as a way to compare the performance of children in one country with that of children in other countries and to understand how education is organized and delivered around the world. Advocates of international studies claim their usefulness lies in the information they provide countries about the strengths and weaknesses of their own educational systems. They believe that identifying the models and practices used in high-performing countries provides lower-performing countries with solutions to their own educational shortcomings (Bloom, 2006; Ben-Simon & Cohen, 2004; Kellaghan, 2004; Keys, 1997).

Many policymakers, educators, and business leaders claim that America's economic future depends directly on the nation's ability to raise its academic standing relative to other countries. Others argue that international rankings are not meaningful. They believe that test score comparisons provide little information about the quality of education in any country. Salzman and Lowell (2008) stated that international test score comparisons "have been stretched far beyond their usefulness." Rotberg (2008) asked the question: "Do you believe that our problems in economic competitiveness would be solved, or even alleviated, if U.S. students answered a few more questions correctly on international assessments?"

Research Services
Office of Assessment, Research, and Data Analysis
1500 Biscayne Boulevard, Suite 225, Miami, Florida 33132
(305) 995-7503 Fax (305) 995-7521

Some researchers have charged that only poor performance gets widely reported by the U.S. media because unfavorable comparisons generate more national interest than "good news" (Hull, 2007; Bracey, 2004; Baker, 1997). Bracey (2004) also contended that the "tendency to accept the results of these studies . . . has been most prominent in those nations that score well."

Concerns About Using International Assessment Results to Evaluate the Performance of Students in the United States

Although the results of international assessments have been used to make broad generalizations about U.S. students' performance, researchers have advised educators to consider the following limitations before drawing conclusions about U.S. students' standing in the global education community.

- Several researchers have claimed that international comparisons are unfair because some countries test only their best students while others test a broad range of students. The samples selected do not always provide a true representation of the varied students, classrooms, and schools in the participating countries (Center for Public Education, 2006; Ben-Simon & Cohen, 2004; Kellaghan, 2004; Prais, 2003; Rotberg, 1998). Prais (2003) stated that this results in an "underlying lack of comparability in the samples chosen to represent each country."

  Furthermore, in most countries, not all of the schools selected to participate in the assessment actually do so. For example, Programme for International Student Assessment (PISA) officials required that a minimum of 85% of sample schools participate in the 2000 assessment. Although most countries met this requirement, PISA score reports included results from several countries with much lower response rates (for example, 61% in the United Kingdom, 56% in the United States, and 27% in The Netherlands) (Prais, 2003).

  Once sample schools are selected, it can be extremely difficult to obtain the full cooperation of all students, parents, and educational staff within each school. As a result, a significant portion of each school's students may not be tested (Ben-Simon & Cohen, 2004; Bradburn & Gilford, 1990). Prais (2003) reported that, when calculating countries' average 2000 PISA scores, the Organisation for Economic Co-operation and Development (OECD) included scores from schools with only a 25% participation rate. He suggested that this cut-off seemed unacceptably low since, if only one quarter of the sampled students in a school participated, they were likely to be the higher-attaining students.

- Some researchers have expressed concern that only countries with the most to gain agree to participate in international assessments (Bloom, 2006; Baker, 1997). Braun and Kanjee (2006) observed that participation in international assessments is "a political decision . . . because of concern about the consequences of poor performance." Bloom (2006) stated that countries furthest from achieving universal education have the least incentive to participate, since they lack resources and capacity, and do not want to publicize their educational failings. In addition, some experts have questioned the manner in which students and schools were sampled by governments with limited funding and the extent to which testing and administration procedures were modified to gain a scoring advantage (Holliday & Holliday, 2003).

- Researchers have claimed that different countries actually compare different populations of schools and students (Holliday and Holliday, 2003). Concerns about differences in the groups compared on international tests include:

  - Studies have shown that when comparing the scores of students at a particular grade level, older students tend to score higher than younger students at the same grade level. For example, if eighth grade students are age 13 in the U.S. but age 14 in Sweden, the Swedish students will be more likely to perform at higher levels (Bracey, 1998a).

  - Studies have also found that when students are the same age but enrolled in different grade levels, students at a higher grade

level will outperform students at a lower grade level. For example, 15 year old students in grade 10 usually score higher than 15 year olds in grade 9 (Bracey, 2004; Prais, 2003).

- Especially at the higher grade levels, some countries have a smaller proportion of students enrolled in school. This introduces bias into comparisons with countries that enroll a larger proportion of students at those grade levels. For example, the smaller proportion of students attending the upper grades in some European countries is likely to be more academically advanced and would therefore receive higher test scores than their peers who are no longer in school (Rotberg, 2008; Prais, 2003).

- Schools in developing countries are more likely to enroll children from higher income families who are usually more academically advantaged (Rotberg, 2008).

- While the U.S. assesses students attending a range of diverse schools, other countries test students attending a narrower group of schools. Some countries test a large proportion of students enrolled in differentiated or streamed schools that separate students based on their academic abilities or career goals. These students may be enrolled in a curriculum that concentrates on math and science, for example, resulting in comparisons between students who have studied a topic such as physics for three years and students in other countries who have studied physics for only one year (Hull, 2007; Bracey, 2004; Holliday & Holliday, 2003; Bracey, 1998a; Rotberg, 1998; Baker, 1997). Prais (2003) reported that the 2000 PISA administration included students attending special schools (slow learners and those with behavioral problems) in some countries but not in others. He found that the exclusion of special schools may have raised countries' average PISA math scores by approximately eight points.

- Some experts have voiced concern that administrative procedures vary too greatly between countries to produce reliable information. Researchers have identified variations in the quality of the administration process across countries, including the caliber of teachers administering the test and the quality of their test administration training; the extent to which instructions are followed (e.g., observing time limits); and the degree of interference in the actual testing process (e.g., assisting or prompting) (Ben-Simon & Cohen, 2004; Keys, 1997; Bradburn & Gilford, 1990).

- The need to create one test in multiple languages introduces a significant source of bias into the development, administration, and scoring of international assessments. It is impossible to construct test items and guarantee equivalency of meanings across languages. Linguists may read a test item expressed in two languages and agree that the wording is identical, but students may interpret the item in very different ways (Ben-Simon & Cohen, 2004; Kellaghan, 2004; Holliday & Holliday, 2003).

  In addition, languages differ in the number of words needed to express an idea and in the number of characters needed to represent a word. The same passage is shorter in English than in Russian and still shorter in Hebrew. Therefore, depending on the language of the student, the amount of space needed for the written test will be shorter or longer and font sizes will differ. These variations mean that students in some countries will need more time to read written text, a confounding factor when tests are administered under strict time limits (Ben-Simon & Cohen, 2004).

- Critics caution that several factors often lead to misinterpretation of international assessment results. First, the overwhelming volume of results reported in most studies, the large amount of methodological details, and the complex relationships among the various measures often make the results unintelligible to almost everyone except measurement experts (Ben-Simon & Cohen, 2004).

  Second, international assessment results are commonly reported as a given country's rank in relation to other participating countries. However, differences in ranking often reflect negligible differences between countries' mean test scores. For example, it is accurate to report that U.S. eighth graders ranked 15 out of 27

countries on a reading test. However, if only three of the countries scored significantly higher than the U.S. and 11 scored statistically the same, the rank does not provide an accurate picture of U.S. students' reading performance. Researchers emphasize that, before drawing conclusions based on rankings, educators must know if the score differences are statistically significant (Salzman & Lowell, 2008; Hull, 2007; The Center for Public Education, 2006; Ben-Simon & Cohen, 2004; Baker, 1997; Bradburn & Gilford, 1990).

- Certain test characteristics may give some countries an advantage over others. For example, a strong emphasis on specific topics may result in higher scores for some countries. Furthermore, particular subject domains may be taught at earlier grade levels in some countries. Finally, when students have more familiarity and experience with a specific item format used in a test, it is likely to have a positive impact on their performance (Ben-Simon & Cohen, 2004; Greaney & Kellaghan, 1996).

- Some experts are concerned that international assessments are not taken as seriously in the U.S. as they are in other countries because the tests are not considered to be high-stakes. Students in countries whose test results have an impact on educational policies and practices, such as retention or graduation, teacher promotions or salary increases, and school budgets, may be more motivated to perform well and more likely to have spent time practicing tasks similar to those that will appear on the test (Ben-Simon & Cohen, 2004; Holliday & Holliday, 2003; Bradburn & Gilford, 1990).

- Studies suggest it may be more beneficial to compare test results obtained at different locations within each participating country. The Organisation for Economic Co-operation and Development (2004) reported that 90 percent of the variance in PISA test scores was observed within, rather than between, countries. Salzman and Lowell (2008) concluded that most of what can be learned about high performance is due to variation in factors within each country. They suggested that it is more effective to transfer best practices across cities and states than to adopt poorly understood practices found in an assortment of small countries around the world.

- A final concern is that international assessment scores are not an accurate measure of school effectiveness. Countries choose different skills, applied to different concepts, to define achievement. In addition, areas such as knowledge of history or mastery of foreign languages are higher priorities in some countries' education systems. Tests measuring students' knowledge and skills in these areas would provide better indicators of their schools' effectiveness than tests focusing on reading, math, and science (Bloom, 2006; The Center for Public Education, 2006; Gradstein & Nikitin, 2004; Greaney & Kellaghan, 1996). Finally, some researchers insist that using only standardized test scores to evaluate education in any country is inappropriate, since they are only one measure of the effectiveness of education (Noack, 1999; Bracey, 1998b). Rotberg (1990) suggested that criteria such as retention and graduation rates, access to higher education for low-income, minority, and disabled students, and the availability of qualified professionals to meet work force requirements can better determine the effectiveness of a country's educational system.

U.S. Performance on International Assessments

The U.S. participates in four international assessments of knowledge and skills (Hull, 2007):

- The Adult Literacy Skills and Lifeskills (ALL) Survey assesses how well adults ages 16-65 apply reading and math skills in life and at work. ALL is conducted by ProLiteracy Worldwide, an international nonprofit organization that sponsors educational services and programs to help adults and their families acquire literacy skills. In 2003, students from six countries participated in the ALL assessment (ProLiteracy Worldwide, 2006).

- The Programme for International Student Assessment (PISA) assesses how well 15 year old students apply knowledge and skills in reading, math, and science. The test moves beyond skills acquired in traditional curricula toward the application of knowledge in everyday tasks. PISA is administered by the Organisation

for Economic Co-operation and Development (OECD), a forum of governments from 30 democratic countries that work together to address the economic, social, and governance challenges of globalization. In 2006, students from 57 countries participated in PISA (OECD, 2008).

- The Progress in International Reading Literacy Study (PIRLS) assesses the reading literacy of fourth grade students. PIRLS is administered by the International Association for the Evaluation of Educational Achievement (IEA), an independent, international cooperative of research institutions and governmental research agencies. In 2006, students from 40 countries participated in PIRLS (IEA, 2007).

- The Trends in International Mathematics and Science Study (TIMSS) assesses how well fourth and eighth grade students understand the math and science concepts taught in school. TIMSS is administered by the International Association for the Evaluation of Educational Achievement (IEA), an independent, international cooperative of research institutions and governmental research agencies. In 2003, students from 46 countries participated in TIMSS, at the fourth grade level, eighth grade level, or both (IEA, 2007).

There is no clear answer as to how well U.S. schools are performing on international assessments compared to other countries. U.S. performance varies considerably depending on the subject area being tested and the age of the students. In general, U.S. students' reading, math, and science scores have been above average at grades four and eight, but have declined as students reach higher grade levels (Baldi et al., 2007; Hull, 2007).

Based on these results, many media reports concluded that American school effectiveness declines as students progress through the grades. A study conducted by the American Institutes of Research (Ginsburg et al., 2005), however, has refuted these claims. When 2003 TIMSS and PISA results were published, most reports stated that 15-year-old U.S. students' scores on PISA declined sharply, compared with higher scores on TIMSS at grades 4 and 8. However, many higher performing countries that participated in PISA and contributed to the lower U.S. rankings were absent from the TIMSS results. The researchers isolated the same group of 12 countries that participated in each of the assessments and reexamined the PISA and TIMSS results. Once the composition of the countries participating in the three assessments was controlled, there was no evidence of a sharp decline on PISA compared with TIMSS, but instead, relatively consistent U.S. performance. The researchers concluded that initially published reports were misleading, due to the different countries the U.S. was compared to on each assessment.

Interpreting International Assessment Results

Because each international assessment compares a different set of countries and measures different subjects, different groups of students, and different types of knowledge, it is difficult to compare the performance of students worldwide. To help interpret the results, Hull (2007) recommended that educators ask the following questions:

- What subjects and grade levels were tested?

- How many countries, and which ones, participated in the assessment? No two international assessments are administered to the same combination of countries. Educators might therefore want to compare countries similar to the U.S., such as *Group of 8* countries (Canada, England, France, Germany, Italy, Japan, Russia, and the United States), because the organization of their educational systems is similar to that of the U.S. and they compete with the U.S. in the economic market.

- Does the test measure the knowledge and skills students obtained in the classroom or their ability to apply this knowledge to real life experiences?

- How are the results reported and what do they really say about students' performance? Many studies report results by ranking countries' relative performance. To make use of the data, educators need to know if the differences between the rankings are statistically significant. Some researchers suggest that it is more informative to analyze international assessment outcomes in terms of the

percentage of students scoring at established proficiency levels (Ben-Simon & Cohen, 2004; Bradburn & Gilford, 1990).

Summary

In conclusion, the results of international assessments can provide useful information, but there are substantial limitations educators should be aware of when comparing the performance of U.S. students to that of students internationally. Some researchers have suggested that it is more effective to transfer best practices across cities and states than to adopt poorly understood practices found in an assortment of small countries around the world. Educators need to understand what each international assessment is actually measuring, who is being assessed, and what the results mean in order to determine how the U.S. stands compared to other countries.

References

Baker, D.P. (1997). Surviving TIMSS: Or, Everything You Blissfully Forgot About International Comparisons. *Phi Delta Kappan, 79*(4), 295-300.

Baldi, S., Jin, Y., Skemer, M., Green, P.J., & Herget, D.(2007). *Highlights from PISA 2006: Performance of U.S. 15-Year-Old Students in Science and Mathematics Literacy in an International Context* (NCES2008-016). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education, Washington, D.C.

Ben-Simon, A., & Cohen, Y. (2004). *International Assessments: Merits and Pitfalls.* Paper presented at the 30th Annual Conference of the International Association for Educational Assessment, Philadelphia, PA, June 2004. Retrieved from http://clickit.ort.org.il/files/upl/224004906/512066429.pdf.

Bloom, D.E. (2006). *Measuring Global Educational Progress.* Cambridge, MA: American Academy of Arts and Sciences. Retrieved from http://www.amacad.org/publications/bloom.pdf.

Bracey, G.W. (1998a). Tinkering with TIMSS. *Phi Delta Kappan, 80*(1), 32-36.

Bracey, G.W. (1998b). The Eighth Bracey Report on the Condition of Public Education. *Phi Delta Kappan, 80*(2), 112-131.

Bracey, G.W. (2004). International Comparisons: Less Than Meets the Eye. *Phi Delta Kappan, 85*(6), 477-478.

Bradburn, N.M., & Gilford, D.M. (1990). *A Framework and Principles for International Comparative Studies in Education.* Washington, D.C.: National Academy Press. Retrieved from http://www.nap.edu/html/framework.

Braun, H., & Kanjee, A. (2006). Using Assessment to Improve Education in Developing Nations. In H. Braun, A. Kanjee, E. Bettinger, & M. Kremer (Eds.), *Improving Education Through Assessment, Innovation, and Evaluation.* Cambridge, MA: American Academy of Arts and Sciences.

The Center for Public Education. (2006). *Criticisms of international assessments: Fact or Fiction?* Retrieved from http://www.centerforpubliceducation.org.

Ginsburg, A., Cooke, G., Leinwand, S., Noell, J., & Pollock, E. (2005). *Reassessing U.S. International Mathematics Performance: New Findings from the 2003 TIMSS and PISA.* Washington, D.C.: American Institutes for Research. Retrieved from http://www.air.org/news/documents/TIMSS_PISA%20math%20study.pdf.

Gradstein, M., & Nikitin, D. (2004). *Educational Expansion: Evidence and Interpretation.* World Bank Policy Research Working Paper 3245. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=610286.

Greaney, V., & Kellaghan, T. (1996). *Monitoring the Learning Outcomes of Education Systems.* Washington, D.C.: The World Bank.

Holliday, W.G., & Holliday, B.W. (2003). Why Using International Comparative Math and Science Achievement Data from TIMSS Is Not Helpful. *Education Forum, 67*(3), 250-257.

Hull, J. (2007). *More Than a Horse Race: A Guide to International Tests of Student Achievement.* The Center for Public Education. Retrieved from http://www.centerforpubliceducation.org.

International Association for the Evaluation of Educational Achievement. (2007). *Mission Statement.* Retrieved from http://www.iea.nl.

Kellaghan, T. (2004). *Public Examinations, National and International Assessments, and Educational Policy.* Retrieved from http://siteresources.worldbank.org/INTAFRREGTOPSEIA/Resources/paper_Kellaghan.pdf.

Keys, W. (1997). What Do International Comparisons Really Tell Us? *International Electronic Journal for Leadership in Learning, 1*(4). Retrieved from http://www.ucalgary.ca/~iejll/volume1/Keysv1n4.html.

Noack, E.G. (1999). Comparing U.S. and German Education: Like Apples and Sauerkraut. *Phi Delta Kappan, 80*(10), 773-776.

Organisation for Economic Co-operation and Development. (2004). *Learning for Tomorrow's World: First Results from PISA 2003.* Paris: Organisation for Economic Co-operation and Development.

Organisation for Economic Co-operation and Development. (2008). *About OECD.* Retrieved from http://www.oecd.org.

Prais, S.J. (2003). Cautions on OECD's Recent Educational Survey (PISA). (2003). *Oxford Review of Education, 29*(2), 139-163.http://www.oecd.org/

ProLiteracy Worldwide. (2006). *The State of Adult Literacy 2006.* Retrieved from http://www.proliteracy.org/downloads/stateoflit06pdf.pdf.

Rotberg, I.C. (1990). I Never Promised You First Place. *Phi Delta Kappan, 72*(4), 296-303.

Rotberg, I.C. (1998). Interpretation of International Test Score Comparisons. *Science, 280*(5366), 1030-1031.

Rotberg, I.C. (2008). Quick Fixes, Test Scores, and the Global Economy. *Education Week, 27*(41), 27, 32.

Salzman, H., & Lowell, L. (2008). Making the Grade. *Nature, 458*(7191), 28-30.

All reports distributed by Research Services can be accessed at http://drs.dadeschools.net under the "Current Publications" menu.