INFORMATION CAPSULE
Research Services

Vol 1306
February 2014

Christie Blazer, Supervisor

EIGHT REASONS WHY INTERNATIONAL ASSESSMENT RESULTS SHOULD BE INTERPRETED WITH CAUTION

At a Glance

Results of international assessments such as the PISA and TIMSS receive a considerable amount of media attention, but scholars have voiced conflicting opinions about the importance of these test results. Some educators and policymakers claim that the unexceptional performance of American students proves that they are not prepared to compete in a global economy and suggests that the U.S. will have serious economic challenges in the future. Others believe that simple country comparisons based on a single assessment score are misleading because they ignore the complexity of test results, encouraging policymakers to pursue inappropriate educational reforms. This Information Capsule summarizes eight reasons why international assessment results should be interpreted with caution.

Educators and policymakers have voiced conflicting opinions about the performance of U.S. students on high-profile international assessments. One group of scholars claims that American students' substandard performance highlights the weaknesses in the U.S.' education system. According to Zhao (2012), there is "another global wave of hand wringing, soul searching, and calls for reform" every time the scores from international assessments are released, with media coverage focusing on how East Asian countries top the rankings. Another group of experts maintains that the results of international assessments are far more complex than the headlines lead us to believe. They insist that conclusions drawn from international test comparisons are oversimplified and misleading and that U.S. performance is not as poor as commonly believed (Buckingham, 2013; Carnoy & Rothstein, 2013a; Cavanagh, 2012; Zhao, 2012; Dalton, 2011; Rotberg, 2011; Loveless, 2010; Boe & Shin, 2005).

The two assessments upon which international comparisons are most frequently based are the Trends in International Mathematics and Science Study (TIMSS) and the Programme for International Student Assessment (PISA). The TIMSS is conducted by the International Association for the Evaluation of Educational Achievement (IEA). The PISA is conducted under the authority of the Organisation for Economic Co-operation and Development (OECD). Scores on the two tests are highly correlated but do not measure the same types of learning. The TIMSS (administered in the fourth and eighth grades) is designed to cover mathematics and science topics that are typically included in the formal school curriculum. In contrast, the PISA tests an age-based sample (15 year olds) to gauge their familiarity with more general concepts and skills related to each subject – the type of knowledge that can be gained both inside and outside of school (Carnoy & Rothstein, 2013a; Loveless, 2013a; Dalton, 2011).

On the most recent administration of the PISA in 2012, U.S. students ranked 24th in reading, 36th in mathematics, and 28th in Science, out of 65 participating countries. Shanghai, China had the highest scores in all three subjects (National Center for Education Statistics, 2012).

Sixty-three countries participated in the most recent administration of the TIMSS in 2011. At fourth grade, U.S. students ranked 11th in mathematics and 7th in science. At eighth grade, U.S. students ranked 9th in mathematics and 10th in science. Singapore led the participating countries in fourth grade mathematics and eighth grade science. Korea was the top scoring country in fourth grade science and eighth grade mathematics (National Center for Education Statistics, 2011).

Researchers recommend that conclusions drawn from international test scores be based on careful analyses of testing databases and consideration of all factors surrounding the administration of each assessment. This Information Capsule summarizes eight reasons why results from international assessments should be interpreted with caution.

1. **Shanghai's high test scores are not representative of China's performance.** Many media announcements claimed that China received the top scores on the 2012 PISA, but test results have only been released for students attending high schools in Shanghai, one of the country's richest cities. Scores were not released for any other province in mainland China. It is true that students in Shanghai outscored all other participating countries in reading, mathematics, and science by an equivalent of almost three years of schooling (Nisen, 2013). However, Shanghai's test scores are not representative of the performance of the country as a whole. For example:
    - Shanghai makes up less than 2% of China's population.
    - Shanghai's per capital gross domestic product is more than double the national average.
    - Approximately 84% of Shanghai high school graduates go to college, compared to 24% nationally.
    - Shanghai has an economically and culturally elite population with systems in place to ensure that students who may perform poorly on tests are not allowed into the public schools. China's hukou system does not allow children from other cities and villages to attend Shanghai high schools. All non-residents of Shanghai must either send their children to private schools (often of low quality) or back to schools in their home cities and villages. Hukous are transferred from generation to generation. The children of migrants, even if they are born in Shanghai, receive their parents' hukou, which their children will also inherit. Consequently, only about 37% of Shanghai's expected population of 15 year olds was tested on the 2012 PISA, representing a privileged subset of students (Loveless, 2014; Hefling, 2013; Loveless, 2013b; Nisen, 2013; Roberts, 2013; Sands, 2013; Rotherham, 2011; Dillon, 2010).

    In China, the PISA was administered in 12 mainland provinces (including Shanghai) and in two special administrative regions - Hong Kong and Macao. However, the Chinese government was allowed to review PISA scores prior to their release and then only allowed the OECD to publish the results from Shanghai, Hong Kong, and Macao (Loveless, 2014; Madda, 2013; Sands, 2013). Tom Loveless (2014) of the Brookings Institution stated, "I have studied and been involved with international testing for a long time, and I know of no situation in which a government has been allowed to limit the region of sampling and to choose which testing data will be released." Since China's students receive educations of greatly varying quality,

Loveless (2013c) concluded that no one will know how well China can perform on an international assessment until it participates, as a country, under the same rules as all other countries.

2. **One test score is not an accurate representation of a country's education system.** When policymakers draw conclusions based on one international test score, they ignore the complexity of a country's education system. Researchers point out that analyses of education systems should consider a variety of educational outcomes, as well as societal and cultural factors that shape school practices within each country. Most experts have therefore concluded that simple comparisons of countries on a single assessment, no matter how interesting they may be, provide little guidance for policy development (Buckingham, 2013; Carnoy & Rothstein, 2013b; Madda, 2013; Cavanagh, 2012; Rotberg, 2011; Rotherham, 2011).

   Complicating the interpretation of international assessment results is the tendency for test publishers to release average national results before underlying databases are made available. Carnoy and Rothstein (2013b) noted, for example, that average national results for the 2011 TIMSS were published well before their underlying databases were released. The researchers stated, "This puzzling procedure ensures that commentators draw quick but ill-informed interpretations and that policymakers can offer inappropriate interpretations of the results without fear of contradiction. Analysis of the database takes time, and headlines from the initial release are sealed in conventional wisdom before scholars can complete more careful study."

3. **The performance of U.S. students on international assessments has remained relatively stable over the years.** Despite frequent media reports to the contrary, American students' skills have not declined over the last decade. Tom Loveless of the Brookings Institution stated that one of the biggest myths of international testing is that the U.S. once led the world and that its students' skills have recently eroded. He noted that the U.S. "was never number one and has never been close to number one on international math tests. . . It is more accurate to say that the United States has always trailed the world on math tests. . . there has been no sharp decline – in either the short or long run" (Loveless, 2010). In fact, the U.S.' average score on the PISA has been relatively stable since the test was first administered in 2000, and it has improved on the TIMSS since 1995 (National Center for Education Statistics, 2012; National Center for Education Statistics, 2011).

   Loveless (2010) pointed to results from the First International Mathematics Study (FIMS). The FIMS, a predecessor of the TIMSS, was administered to 13-year-old students in 1964. The U.S. ranked 11th out of 12 participating countries, below Australia, Belgium, England, Finland, France, Germany, Israel, Japan, Netherlands, and Scotland. Only Sweden scored lower.

4. **Rankings and average scores are entirely dependent upon which countries participate in each administration of an international assessment.** The number of countries participating in international assessments has increased steadily over the years. For example, the number of countries administering the PISA doubled from 32 in 2000 to 65 in 2012. The biggest impact on rankings has been the addition of East Asian countries, which now dominate the top rankings on all international assessments. Researchers point out that previously high performing countries slide in the rankings, even if their scores remain stable, when new countries that score in the top tiers of the

performance distribution participate in the testing (Buckingham, 2013; Cavanagh, 2012). Andreas Schleicher, Deputy Director for Education and Special Advisor on Education Policy at the OECD, stated that from a statistical standpoint, "there is no decline on any measure that we have for the United States." He added that "the rate of improvement in other countries, in terms of getting more people into school and educating them well, is steeper" (quoted in Cavanagh, 2012).

Another factor that biases rankings on international assessments is that the countries choosing to participate in each test vary from year to year, depending on which governments want to grant access to their schools and finance the testing. Koretz (2009) noted, "So you'll see newspapers and sometimes our government reporting whether or not the United States scored above or below the international average. But there is no international average, other than the average of the countries that happened to participate that time."

5. **Researchers have not found a strong relationship between international assessment scores and countries' economic growth.** Some analysts have tried to link student performance on international assessments to their nation's economic growth, but most experts agree that the relationship between the educational and economic performance of countries, as measured by tests like the TIMSS and PISA, is tenuous at best (Cavanagh, 2012; Zhao, 2012). Rotberg (2011) concluded that test score rankings cannot be used to accurately predict economic trends. He noted that other variables, such as tax rates, health care and retirement costs, incentives for innovation, intellectual property enforcement, and natural resources are much more predictive of a country's economic competitiveness than mathematics and science test scores. Hal Salzman, an economist at Rutgers University, stated, "If the reason we're concerned about education is economic competition, it's worth noting that a large portion of those high-ranking countries are economic train wrecks" (quoted in Cavanagh, 2012).

6. **Sampling procedures may create opportunities for countries to manipulate the selection of schools and students.** Several researchers have criticized the procedures the OECD uses to select PISA samples. The OECD uses an independent research firm to draw a two-stage stratified random sample. First, schools are randomly selected and then students within the selected schools are randomly chosen for testing. Two potential replacements are selected for each school. If the originally selected schools do not participate in the testing, countries instead test students at the replacement schools. Replacement schools may represent up to 35% of the sampling frame (Carnoy & Rothstein, 2013a; Sands, 2013; Rotberg, 2011).

   Some researchers have suggested that this sampling process allows countries to manipulate which students and schools are tested. For example, countries might exclude academically weaker students in the selected schools. In addition, countries might choose to substitute originally selected schools if they are academically weak and administer the test at higher-performing replacement schools. Furthermore, in order to be included in the PISA sample, students must be enrolled in school, so samples are not representative of the 15 year old population in countries or areas with high dropout rates. Researchers also point to the sampling bias in China, where provinces were not randomly selected at all. Instead, the PISA was administered only in provinces that were approved for participation by the Chinese government (Loveless, 2013c; Roberts, 2013; Sands, 2013; Dillon, 2010).

7. **The U.S. has more low-income students than many other participating countries.** Researchers have found an achievement gap based on income level in every nation. In other words, students from the most affluent homes receive on average the highest test scores and students from the poorest homes receive on average the lowest test scores, regardless of their home country (Ravitch, 2013; Riddile, 2010; Ashcraft, 2009).

   Economic Policy Institute researchers Carnoy and Rothstein (2013a) analyzed 2009 PISA data, disaggregated by students' income levels (based on free or reduced price lunch status). They found that part of the reason the U.S. received lower average PISA scores was because it had more low-income students than many of the countries with which it was compared. Carnoy and Rothstein reported that U.S. PISA scores were even further depressed because the sampling procedures used by the OECD resulted in the inclusion of a disproportionately large percentage of U.S. students attending low-income schools. (The OECD claims that the PISA sample accurately reflected the percentage of low-income schools within the U.S.) (Carnoy & Rothstein, 2013a; Carnoy & Rothstein, 2013c; Education Week, 2013).

   Carnoy and Rothstein (2013a) compared U.S. 2009 PISA scores with scores from the three top-scoring countries (Canada, Finland, and Korea) and from three other countries with which the U.S. is frequently compared (France, Germany, and the United Kingdom). The authors reached the following conclusions:
   - In every country, low-income students received lower scores than higher-income students. In the U.S., average test scores were lower partly because there were a greater proportion of low-income test takers.
   - PISA over-sampled low-income U.S. students who attended schools with very high proportions of similarly disadvantaged students. While 40% of the U.S. PISA sample was drawn from schools where one-half or more of students were eligible for free or reduced price lunch, only 32% of students nationwide attended such schools.
   - Re-estimated U.S. PISA scores that adjusted for the oversampling of disadvantaged students in high poverty schools raised the U.S. PISA ranking from 14th to 6th in reading, and from 25th to 13th in mathematics.
   - The achievement gap between low- and high-income students was actually smaller in the U.S. than it was in France, Germany, and the United Kingdom, and not much larger than the gap in the highest scoring countries (Canada, Finland, and Korea).
   - An examination of trends over the last decade on multiple administrations of both the TIMSS and PISA found that the performance of low-income U.S. students has been improving over time, while the performance of similarly disadvantaged students in the six comparison countries has been falling.

8. **The U.S. has more foreign born citizens and second language learners than most of the highest scoring countries.** Researchers argue that city-states, such as Singapore and Hong Kong, and many of the countries outranking the U.S. on international assessments have smaller immigrant populations and fewer second language learners (Hefling, 2013; Dalton, 2011; Pellissier, 2010; Ashcraft, 2009).

   For example, over 40 million foreign-born people, representing 13% of the population, resided in the U.S. in 2013. But in Finland, only 5% of the population (less than 300,000 people) claimed to have a foreign background (Dews, 2013; Hefling, 2013; Statistics Finland, 2013; Tanner, 2011). The OECD (2011) reported that on average across

countries participating in the 2009 PISA, native students outperformed their immigrant peers by 43 points. (The gap was reduced to 27 score points – still a performance gap of over one-half of a school year – when controlling for socioeconomic status.)

Summary

The results of international assessments are far more complex than media headlines suggest. Researchers urge educators and policymakers to base their conclusions on careful analyses of testing databases and consideration of all factors surrounding the administration of each particular assessment, especially the particular population of students included in the testing. This Information Capsule summarized eight reasons why results from international assessments should be interpreted with caution.

1. Shanghai's high test scores are not representative of China's performance.
2. One test score is not an accurate representation of a country's education system.
3. The performance of U.S. students on international assessments has remained relatively stable over the years, contrary to media reports claiming that American students' skills have declined.
4. Rankings and average scores are entirely dependent upon which countries participate in each administration of an international assessment.
5. Researchers have not found a strong relationship between international assessment scores and countries' economic growth.
6. Sampling procedures may create opportunities for countries to manipulate the selection of schools and students.
7. The U.S. has more low-income students than many other participating countries.
8. The U.S. has more foreign born citizens and second language learners than most of the highest scoring countries.

References

Ashcraft, C. (2009). *Comparing U.S. K-12 Students' Math and Science Performance Internationally: What Are the Facts, What Do They Mean for Educational Reform, and How Do I Talk Effectively About the Issues?* National Center for Women & Information Technology, Boulder, CO. Retrieved from http://www.ncwit.org/resources/comparing-us-k-12-students-math-and-science-performance-internationally-what-are-facts-1.

Boe, E.E., & Shin, S. (2005). Is the United States Really Losing the International Horse Race in Academic Achievement? *Phi Delta Kappan, 86*(9), 688-695.

Buckingham, J. (2013). Don't Panic About PISA. *Australian Broadcasting Corporation,* December 4, 2013. Retrieved from http://www.abc.net.au/news/2013-12-04/buckingham-pisa-panic/5133364.

Carnoy, M., & Rothstein, R. (2013a). *What Do International Tests Really Show About U.S. Student Performance?* Economic Policy Institute, Washington, DC. Retrieved from http://s2.epi.org/files/2013/EPI-What-do-international-tests-really-show-about-US-student performance.pdf.

Carnoy, M., & Rothstein, R. (2013b). *International Tests Show Achievement Gaps in All Countries, with Big Gains for U.S. Disadvantaged Students.* Economic Policy Institute, Washington, DC. Retrieved from http://www.epi.org/blog/international-tests-achievement-gaps-gains-american-students/.

Carnoy, M., & Rothstein, R. (2013c). *Response from Martin Carnoy and Richard Rothstein to OECD/PISA Comments (by Andreas Schleicher, OECD Deputy Director for Education and Special Advisor on Education Policy to the OECD's Secretary-General, January 14, 2013) Regarding Our Report, "What Do International Tests Really Show About American Student Performance?"* Economic Policy Institute, Washington, DC. Retrieved from http://www.epi.org/files/2013/EPI-Carnoy-Rothstein-Resp-to-Schleicher.pdf.

Cavanagh, S. (2012). U.S. Education Pressured by International Comparisons. *Education Week, 31*(16), 6-10.

Dalton, B. (2011). *US Educational Achievement on International Assessments: The Role of Race and Ethnicity.* RTI International, Research Triangle Park, NC. ERIC Document Reproduction Service No. ED535873.

Dews, F. (2013). *What Percentage of U.S. Population is Foreign Born?* Brookings Institution, Washington, DC. Retrieved from http://www.brookings.edu/blogs/brookings-now/posts/2013/09/what-percentage-us-population-foreign-born#.

Dillon, S. (2010). Top Test Scores From Shanghai Stun Educators. *The New York Times,* December 7, 2010.

Education Week. (2013). *OECD/PISA's Response to the Paper "What do International Tests Really Show about American Student Performance" by Martin Carnoy and Richard Rothstein.* Retrieved from http://www.edweek.org/media/globalachievementsstudy-blog.pdf.

Hefling, K. (2013). *Asian Nations Dominate International Test.* Retrieved from http://news.yahoo.com/asian-nations-dominate-international-test-100159386.html.

Koretz, D. (2009). Measure for Measures: What Do Standardized Tests Really Tell Us About Students and Schools? *Harvard Graduate School of Education Usable Knowledge.* Retrieved from http://www.uknow.gse.harvard.edu/decisions/video-DD315-608-1.html.

Loveless, T. (2010). *The 2010 Brown Center Report on American Education: How Well Are American Students Learning?* Brookings Institution, Washington, DC. Retrieved from http://www.brookings.edu/~/media/research/files/reports/2011/2/07 education loveless/0207_education_loveless.pdf.

Loveless, T. (2013a). *International Tests Are Not All the Same.* The Brookings Institution, Washington, DC. Retrieved from http://www.brookings.edu/blogs/brown-center-chalkboard/posts/2013/01/09-timss-pisa-loveless.

Loveless, T. (2013b). *Attention OECD-PISA: Your Silence on China is Wrong.* The Brookings Institution, Washington, DC. Retrieved from http://www.brookings.edu/blogs/brown-center-chalkboard/posts/2013/12/11-shanghai-pisa-scores-wrong-loveless.

Loveless, T. (2013c). *PISA's China Problem.* The Brookings Institution, Washington, DC. Retrieved from http://www.brookings.edu/blogs/brown-center-chalkboard/posts/2013/10/09-pisa-china-problem-loveless.

Loveless, T. (2014). *PISA's China Problem Continues: A Response to Schleicher, Zhang, and Tucker.* Brookings Institution, Washington, DC. Retrieved from http://www.brookings.edu/blogs/brown-center-chalkboard/posts/2014/01/08-shanghai-pisa-loveless.

Madda, M.J. (2013). *How I Learned to Stop Worrying About the PISA.* Retrieved from https://www.edsurge.com/2013-12-03-how-i-learned-to-stop-worrying-about-the-pisa.

National Center for Education Statistics. (2011). *TIMSS Results.* Retrieved from http://nces.ed.gov/timss/results11_math11.asp.

National Center for Education Statistics. (2012). *Selected Findings from PISA 2012.* Retrieved from http://nces.ed.gov/surveys/pisa/pisa2012/.

Nisen, M. (2013). Why Shanghai's Amazing Test Scores Are "Almost Meaningless." *Business Insider,* December 3, 2013. Retrieved from http://www.businessinsider.com/shanghai-pisa-test-scores-2013-12.

Organisation for Economic Co-operation and Development. (2011). *How are School Systems Adapting to Increasing Numbers of Immigrant Students?* Retrieved from http://www.oecd.org/pisa/pisaproducts/pisainfocus/49264831.pdf.

Pellissier, H. (2010). *The Finnish Miracle.* Retrieved from http://www.greatschools.org/students/2453-finland-education.gs.

Ravitch, D. (2013). *Good News: Major Re-Analysis of International Tests.* Retrieved from http://dianeravitch.net/2013/01/25/good-news-major-re-analysis-of-international-tests/.

Riddile, M. (2010). *PISA: It's Poverty Not Stupid.* National Association of Secondary School Principals, Reston, VA. Retrieved from http://nasspblogs.org/principaldifference/2010/12/pisa_its_poverty_not_stupid_1.html.

Roberts, D. (2013). Want to Look Great on Global Education Surveys? Test Only the Top Students. *Bloomberg Businessweek,* December 4, 2013. Retrieved from http://www.businessweek.com/articles/2013-12-04/want-to-look-great-on-global-education-surveys-only-test-the-top-students.

Rotberg, I.C. (2011). International Test Scores, Irrelevant Policies. *Education Week, 31*(3), 32.

Rotherham, A.J. (2011). Shanghai Surprise: Don't Sweat Global Test Data. *Time,* January 20, 2011.

Sands, G. (2013). Are Chinese Students Smarter or is Testing System Rigged in Their Favour? *South China Morning Post,* December 18, 2013. Retrieved from http://www.scmp.com/comment/insight-opinion/article/1385217/are-chinese-students-smarter-or-testing-system-rigged-their.

Statistics Finland. (2013). *Nearly Every Tenth Person Aged 25 to 34 of Foreign Origin.* Retrieved from http://www.stat.fi/til/vaerak/2012/01/vaerak_2012_01_2013-09-27_tie_001_en.html.

Tanner, A. (2011). *Finland's Balancing Act: The Labor Market, Humanitarian Relief, and Immigrant Integration.* Retrieved from http://www.migrationinformation.org/feature/display.cfm?ID=825.

Zhao, Y. (2012). *Numbers Can Lie: What TIMSS and PISA Truly Tell Us, If Anything?* Retrieved from http://zhaolearning.com/2012/12/11/numbers-can-lie-what-timss-and-pisa-truly-tell-us-if-anything/.